How Violence Affects Protests

Zachary C. Steinert-Threlkeld[1], Alexander Chan[2], and Jungseock Joo[3]

[1]University of California, Los Angeles. Email: zst@luskin.ucla.edu

[2]University of California, Los Angeles. Email: alexander.chan@ucla.edu

[3]University of California, Los Angeles. Email: jjoo@comm.ucla.edu

ABSTRACT

A key determinant of whether social movements achieve their policy goal is how many people protest. How many people protest is in turn partially determined by violence from protesters and state agents. Previous work finds mixed results for violence. This paper reconciles the mixed results for violence by distinguishing between its timing, source, and severity: low levels of state repression increase protest size while high levels decrease it, conditional on preventative repression failing. Protester violence is always associated with fewer subsequent protesters. These claims are substantiated by applying deep learning techniques to geolocated protest images shared on social media. Across more than 4,300 observations of twenty-four cities from five countries, we find that protester violence is always associated with subsequently smaller protests, while low (high) levels of state violence correlate with increased (decreased) protest size. The paper ends with a discussion of situations in which to prefer images or text for studying protests; ethical concerns; and improving data collection in order to apply the analysis to poorer or less populous environments.

1 INTRODUCTION

Protests are more likely to shift policy the larger they are (Chenoweth and Stephan, 2011; Fassiotto and Soule, 2017; Gause, 2018). In turn, their size is affected by the level and source (protesters or state agents) of violence at a protest. While protester violence has consistently been found to correlate with smaller protests (Stephan and Chenoweth, 2008; Feinberga et al., 2017; Murdie and Purser, 2017), the protest-repression literature consistently finds inconsistent results (Davenport, 2007). This paper suggests a solution to the protest-repression puzzle rooted in the timing and level of state violence.

Pre-emptive state repression decreases protests (Sullivan, 2016). Once protests start, however, the effect of state violence depends on its severity. We develop and test an argument that low levels of state violence lead to larger protests, while high levels decrease it; protester violence leads to smaller protests.[1] The importance of the severity of state repression may explain varying effects that the literature identifies for state repression (Carey, 2006; Ritter, 2013), though it contradicts the backlash hypothesis (Francisco, 2004).

These theoretical expectations are tested using protest images shared in geolocated tweets. This measurement occurs using three convolutional neural networks (CNN). We first develop a CNN to recognize protest images, and we have verified that this classifier outperforms Google Vision on our images (see Figure A8). From a corpus of six years of geolocated tweets, we identify 55.6 million from protest waves across fourteen countries, 5.4 million of which contain an image. The protest detection classifier identifies over 115,000 of these images as very likely to contain a protest, and we build a second CNN to measure protester and state violence as well as the presence of fire or police officers. Figure A8 also shows that this scene classifier identifies police better than Google Vision. This scene classifier is complemented with the third CNN, a face classifier. This classifier counts the number of faces per photo and estimates the race, gender, and age of each face, allowing us to control for well-known correlations between these demographic features and protest participation. While many off the shelf face classifiers exist, only one codes race precisely enough (Kärkkäinen

[1] Violence should be conceived of as "perceived violence", a point to which we return later.

and Joo, 2019). Sections 3 and S1 detail how we built these models and verify their output.

Section 4 discusses the resulting data, concerns about selection bias, and validates the dependent variable. It shows that users who tweet geolocated protest images are likely to be more representative of normal Twitter users than those who tweet geolocated non-protest images, and there are strong reasons to expect social media to be no more biased than newspapers in covering protests. That section also shows that using images to measure protest size correlates with estimates of protest size from newspapers and records changing activity that matches events as reported in newspapers.

Section 5 presents the main results. It also shows three sets of robust checks, two explicitly designed to account for possible bias. Restricting results to users of moderate popularity; non-verified users; non-bots; and tweets in a country's *lingua franca* do not change inferences. Deduplicating images also does not change inferences. To the extent we find evidence for bias, it is against the main results: dropping bots and restricting tweets based on language both produce better model fit than the full datasets.

Finally, Section 6 concludes, discussing why images, instead of text, are necessary for this project; the ethical concerns raised by computer vision approaches, especially in the context of contentious politics; and why the results presented in this paper should be considered a lower bound on what these techniques can achieve.

2 PROTEST DYNAMICS

Protesters aim to convince bystanders to mobilize, increasing pressure for policy change (**?**). The state works to convince bystanders to remain on the sidelines and existing protesters to disengage. Protesters and the state each choose amounts of violence to employ. Protester violence should always lead to smaller protests, while state violence will have differential effects depending on its timing and severity.

2.1 Large Protests

Three assumptions lead to the conclusion that large protests are more likely to change policy than small ones. If (1) the purpose of a protest is to convince political leaders to change a policy, (2) a leader cares about the median voter (Downs, 1957) or his or her winning coalition exhibits

some response to the median person (Bueno de Mesquita et al., 2003), and (3) a large protest's policy preference is closer to the median individual than a small one's, then a large protest is more likely to change policy than a small one.

This argument can also be obtained without assuming a leader aims for the median individual's policy preference. If a leader only desires to stay in power and a large protest means the probability of remaining in power is lower than the leader previously believed, a large protest is still more likely to lead to policy change than a small one. While a large protest is not necessarily successful, all successful protests are large.

Protest size is important regardless of a country's political institutions. In democracies, voting is the most common method of policy change. The aggregation of preferences through defined rules, and the willingness of those in power to heed the result, has many advantages. It is a low-cost endeavor for participants, as the only costs are transaction and opportunity. Voting occurs infrequently, however, and is a blunt method of feedback because it collapses all political dimensions into one. Protests, however, can occur at any time and usually have a clear policy goal (Battaglini, 2017).

In authoritarian regimes, however, voting is a less significant act. In countries where policy feedback comes from an insider population drawn from the larger populace (Bueno de Mesquita et al., 2003), those belonging to the outsider population provide policy feedback through protest (or rebellion). While protest is unlikely to change an autocrat's policy, it nonetheless provides a key signal of discontent to which a government can respond (Bratton and Walle, 1992). This signal is especially pertinent if opinion polling is unreliable (Robertson, 2007) or the media are not free (Qin et al., 2017).

The importance of large protests has not escaped notice. Lohmann (1994a) argues that unprecedented numbers of people rallying in the German Democratic Republic in the beginning of 1989 were a key reason the protest movement grew and eventually toppled Erich Honecker's government. The gradual growth of protest size in Iran in 1979 also made it increasingly difficult for the Shah to remain in power (Rasler, 1996). Kuran (1989)'s canonical model of bandwagoning implicitly

means a revolution follows large protests. Understanding the threat posed by large crowds, regimes often raise the cost of protesting by killing protesters, yet killing protesters has an indeterminate effect on the size of subsequent protests (Francisco, 2004). Indeed, if there is a law like regularity to the study of protest mobilization, it is that "size matters" (Biggs, 2016).

The importance of size applies to social movements as well, of which protests are a tactic they can employ (Tilly and Wood, 2012). Large social movements are more likely to lead to policy change than small ones for several reasons. Because large social movements tend to be nonviolent, they increase the domestic and international cost of repression, especially when movements maintain their own media (Sutton et al., 2014). They decrease the cost of participation, making individuals more likely to join, and making them more likely to join as movements grow (DeNardo, 1985). It also increases the probability that individuals within the winning coalition defect, making it more difficult for the state to continue repression (Goldstone, 2001). For a fuller exposition of the importance of size for social movements, see Chenoweth and Stephan (2011).

2.2 Violence

Though scholars understand the importance of large protests, less is understood about why protests become large, and most existing work is qualitative, cross-sectional, or focuses on structural variables. For example, Biggs (2003) argues, in explaining American protests in 1886, for a positive feedback loop but does not specify when an initial protest is more likely to generate that process. Large protests in one country may occur because large protests in a similar country succeeded, but contagion does not explain why the initiating country experienced large protests (Weyland, 2012). The structure of built environments may also encourage protest participation: one reason initial marches to Tiananmen Square succeeded is because universities in Beijing are in the same neighborhood and have internally dense configurations, encouraging mobilization both within and between campuses (Zhao, 1998). The occurrence of electoral fraud is also a common source of large protest events (Tucker, 2007).

This paper focuses on mechanisms affecting protest size *once a protest starts*. The only kind of repression that can occur in this situation is overt repression, often called "protest policing"

(Earl, 2003; Davenport and Soule, 2009). This concept refers to repressive behaviors that occur during a protest, such as blocking roads, impeding pedestrian movement, arresting protesters, and using subfatal weapons such as tear gas, water cannons, or sound guns. Protest policing contrasts with preventative repression, such as the targeting of dissident organizations or arresting particular individuals (Sullivan, 2016).

Protester Violence

If violent protests are more costly to the individual than nonviolent ones, regardless of the source, then violence should decrease protest size (Moore, 1995). Empirically, however, there appears to be differential effects based on the source of the violence.

When violence originates from protesters, it should always decrease the size of protests because it decreases the number of people to which protest appeals, and it increases the cost of protesting to the remaining bystanders who may protest.

One method by which bystanders determine whether or not to join a protest is to compare protesters' ideological distance to their own (Lohmann, 1993). Since most individuals do not support violence or receive consumption value from it (Feinberga et al., 2017), violence originating from protesters signals that protesters, and therefore the policy changes for which they agitate, are far from mainstream. Survey research has found that the more activists differ from the population they try to mobilize, the less likely individuals are to protest (Bashir et al., 2013). Being far from the mainstream, bystanders continue to stand by because the new policy the violent protesters seek is inferred to not be beneficial.

Protester violence decreases the likelihood of regime defections, decreasing the number of non-protesters available to mobilize. Peaceful protest convinces regime agents of their physical safety should they defect, increasing the probability that police, members of the armed forces, or legislators, for example, switch allegiances (Stephan and Chenoweth, 2008). Violent protests, however, induce fear in these agents that they will meet the same fate if they do not remain loyal. Violence therefore reduces the pool of those willing to protest, making the state stronger than an equivalent peaceful protest.

Since protester violence decreases the legitimacy of protests, the state can pursue high levels of repression and face less risk of backlash. Peaceful protests enjoy high domestic and international legitimacy, so state violence against them risks generating a backlash that increases subsequent protests' size (Francisco, 2004). But since violent protesters can be framed as rioters, terrorists, or foreign agitators (Benford and Snow, 2000), bystanders are more supportive of repressing violent protests than nonviolent ones. Survey work across eighteen countries finds that violent protests decrease future support for the peaceful right to protest (Murdie and Purser, 2017). For the same reasons, the state is also less likely to receive international sanction when repressing violent protests.

The converse of these arguments is that protester non-violence increases the probability that a protest grows in size, especially when states repress. Because non-violence increases the legitimacy of protests, it decreases the probability that a state represses, as the state will pay large reputation costs. The lower probability of repression induces more bystanders to mobilize, generating a positive feedback loop (Lohmann, 1994b). In Morocco, for example, attempts to repress non-violent protesters at the start of the Arab Spring led to larger crowd sizes (Lawrence, 2016), and government violence in Tunisia did not prevent the spread of those protests.

Since protester violence alienates bystanders, increases the resolve of state agents, and invites high levels of state repression, we expect that:

H1: There should exist a negative relationship between protester violence and the subsequent size of a protest.

This hypothesis extends earlier work that finds the same relationship at the movement level, using a movement's reported maximum participation rate. As far as we are aware, existing work on protester violence and outcomes is cross-sectional (Stephan and Chenoweth, 2008; Celestino and Gleditsch, 2013; Chenoweth and Schock, 2015) or focused on its interaction with state tactics (Shellman et al., 2013). It is therefore unclear if protester violence decreases participation, less participation causes protesters to result to violence, or a smaller movements results from another feature. By developing a logic for protester violence and individual participation, we directly link

these two and explain how the former should affect the latter's fluctuation.

State Violence

While the negative relationship between protester violence and movement success is a regular finding, the literature on state repression and protest has not found consistent effects. In Peru and Sri Lanka, repression decreased subsequent protests (Moore, 2000). The same has been found in West Germany (Koopmans, 1993), South Africa (Olzak et al., 2003), Iran in the short-term (Rasler, 1996), and the Middle East and North Africa during the Arab Spring (Steinert-Threlkeld, 2017). On the other hand, repression may have increased protest in West Germany and Ireland (Francisco, 1996) and Iran with a six-week lag (Rasler, 1996), and many cross-national studies find repression increases protest (Gurr and Moore, 1997; Davenport and Armstrong II, 2004; Francisco, 2004; Hess and Martin, 2006). In can also increase protest based on the emotional reaction of individuals connected to those targeted (Siegel, 2011; Pearlman, 2013). On the third hand, there is sometimes no correlation between repression and protest levels (Gupta et al., 1993; Ritter, 2013; Ritter and Conrad, 2016).

These contradictory findings are resolved by considering the timing of repression and the severity of it. When mobilization is the result of social movement organizations' planning, repression focusing on those organizations should decrease protest size (Sullivan, 2016). This preemptive repression attacks the infrastructure of protests, making it harder for them to occur, much less grow (Danneman and Ritter, 2013; Sutton et al., 2014). This line of reasoning then argues that repression of protests as they occur leads to backlash (Sullivan, 2016). Repression of protests as they occur, commonly called protest policing (Della Porta and Reiter, 1998; Davenport and Soule, 2009; Earl et al., 2013), leads to the differential effects discussed earlier.

Light repression will generate backlash for two reasons. First, they may signal that the cost of protesting is lower than bystanders believed. Now aware that protesting is a net positive, bystanders join those already protesting. Second, repression can generate emotions such as anger, joy, or pride. Acting on these emotions provides intrinsic benefit to the former bystander, regardless of instrumental calculations (Pearlman, 2013). Incorporating emotions into theories does not require

avoiding rationality assumptions, as protesting in anger at repression can be individually rational (Siegel, 2011).

Severe repression, however, should lead to smaller protests, for similar reasons. Severe repression may signal that state actors are more resolved than protesters expected. Facing a higher cost to protest, protesters become bystanders. Severe repression also generates fear, sadness, and shame, causing protesters to deactivate and bystanders to remain where they are (Pearlman, 2013). This emotional effect has also received recent support in a series of lab-in-the-field experiments in Zimbabwe (Young, 2019).[2]

For an earlier exposition of a similar argument, see Gurr (1970). For a formal derivation of this relationship, see DeNardo (1985). Observational studies which distinguish types of repression by the cost they impose also find that severe repression decreases mobilization (Muller, 1985; Khawaja, 1993). In other words, the contradictory effects may be due more to measurement error than theoretical inconsistencies. Since it appears that apparently contradictory effects of repression are resolved by stipulating the severity of repression, conditional on observing protest, we expect that:

H2: There is an n*-shaped relationship between between state repression and the subsequent size of a protest.*

H2 should apply in democracies and autocracies. For example, the Occupy Wall Street movement in the United States did not grow large until New York City police arrested over 700 participants, in a manner many perceived as unjust, marching on the Brooklyn Bridge. The movement waned six weeks later, in the middle of November 2011, once local police forcibly dismantled protesters' main encampment at Zuccotti Park and forbade them from spending the night (White, 2016). in a m This effect should occur in democracies and autocracies. In Egypt, the protests starting on January 25^{th} were met with initial state resistance and some casualties; 18 days later,

[2]Francisco (2004) finds that state massacres increase mobilization. This result is due to an expansive definition of mobilization: the majority of the backlash events are substitutes for mobilization because they are harder to repress (Moore, 2000). Our focus is on mobilized protesters, not all forms of mobilization.

the Armed Forces forced President Hosni Mubarak to abdicate. Two years later, the Armed Forces launched a coup against the elected president, Mohamed Morsi. Large pro-Morsi protests erupted and continued for six weeks. The Armed Forces initial attempts to demobilize them caused them to grow in size; morning massacres on August 14^{th} at the two main encampments killed at least 1,000 protesters and injured even more (Shakir, 2014).

H2 at first appears inconsistent with the backlash hypothesis substantiated in Francisco (1995), Francisco (1996), and Francisco (2004). It is not. That body of work argues against an "inverse-u" relationship between state repression and protest. Instead, evidence of backlash is found: when states engage in severe repression, the response is more collective action. That work, however, broadens protest to include other forms of collective action such as strikes, building occupations, or guerrilla action. Moreover, the substitution that does occur most often does not occur the day immediately following the repression. In other words, when the state meets protesters with severe repression, they initially reduce their protest; after some delay, they backlash by substituting away from direct confrontation with the state.

The argument put forth in this subsection is that severe repression decreases protest. It does not make a claim about whether other types of dissent increase. Works such as Francisco (1995), Francisco (1996), and Francisco (2004) define backlash in a more encompassing method than we do. This different definition is why they initially appear to have different expectations about, and different results for, backlash. H2 is not inconsistent with that backlash hypothesis because it is focused on a narrower window and action repertoire.

3 AUTOMATED CODING OF SOCIAL MEDIA IMAGES

To test the aforementioned hypotheses, we need to measure many variables, such as protester age or gender. For this task, we turn to images shared with geolocated tweets, explanation of which is provided in Sections 4 and S4 through S6. Since the target dataset is very large, we develop three automated classifiers based on convolutional neural networks to automatically code the variables of interest: one to identify protest images, a scene classifier to extract data (primarily about violence)

from protest images, and a face classifier to generate cleavage and size information.[3] Table 1 provides an overview of the steps required in this pipeline, and the rest of this section provides a very brief introduction. The rest of this section describes the two classifiers we developed and one already existing one we used. For a high-level overview of how convolutional neural networks works, see Section S1. For validation of the classifiers' results, see Section S1.3 as well as Section 4.3.

Table 1. Protest Data Pipeline

Steps	Input	Source	Output
	Collecting Images for Training Set		
1. Image search	Keywords	Google	100,000 images
2. Train a protest image classifier	Images from Step 1	Self	Initial CNN
3. Protest images from Twitter corpus	Model from Step 2	Twitter	40,764 images
	Developing Protest and Scene Classifier		
4. Manual annotation	Images from Step 3	Amazon Mechanical Turk	13 ground-truth labels
5. Train a CNN	Training data from Step 4	Self	Protest and scene classifier
	Face Attribute Classification		
6. Face classification	-	(Kärkkäinen and Joo, 2019)	Gender, age, and race estimates

3.1 Image Collection for Protest Classification by Weakly-supervised Learning

Step 1. As typically done in supervised machine learning, our approach in model development begins with collecting training data: images and target classification labels. Images in a training set should exhibit diverse visual traits of protest events and also include a range of negative (non-protest) images such that the trained classifier generalizes well to unseen images. In addition, it is desirable that the set also contains many difficult images, hard negatives, i.e. non-protest images which look like protest scenes, to make the classifier more robust.

The efficiency of manual annotation to collect target labels is another important consideration. For example, sampling general images and providing them to annotators would create a training set

[3]The first two classifiers are in fact partially combined in implementation such that one integrated classifier can generate two sets of outputs, although they differ conceptually. This is called multi-task learning (Girshick, 2015). We still discuss two classifiers separately because they are trained on different data and used in different steps.

of mostly non-protest images. This approach is not cost effective. Therefore, we take a combination of weakly-supervised and supervised learning. In weakly-supervised learning, the ground-truth labels on the target variable are not directly available but can be inferred from other variables (Bergamo and Torresani, 2010). For instance, we can use any online image search service to query images with a particular keyword (e.g., "protest"), and this step will furnish a large quantity of relevant images. While this sample set will contain some noisy data, it is still useful to train a rough initial model which can be used to fetch better samples. These samples can be manually annotated as in typical supervised learning.

Specifically, we first collected about 10,000 protest images from Google Image Search by using manually selected keywords such as "protest," "riot," "Black Lives Matter," "Venezuela Protest," "Hong Kong protest" and many others, as well 90,000 non-protest, hard-negative images by using keywords including "concert," "stadium," or "airport crowd." These negative examples are called hard-negatives because they look similar to protest images (e.g., crowded), and classifiers can easily misassign their labels. Since these images are simply outputs of search queries, their assigned labels are not accurate. For example, the query of "protest" may return a few photographs of politicians. However, we did not verify the correct classification labels of these images because the main purpose of this first step is to train a rough classifier with the assumption that the majority of labels are still correct.

Step 2. Using these data, we trained a convolutional neural network (CNN) whose only output denotes whether an image captures a protest event or not. We then applied this classifier to geolocated images from Twitter and obtained the classification scores. Each score can be considered as the confidence about the output, the probability of the input image containing protesters. Section S1.1 provides detail of how CNNs work and the specific architecture of this paper's, and Joo and Steinert-Threlkeld (2018) provides a detailed explanation of their relevance to political science.[4]

Step 3. Twitter provides tweets in real time through its streaming application programming interface (API). Since late 2013, one of the authors has used this interface to collect tweets with

[4]See as well (Cantu) for an application of this methodology to vote fraud detection.

longitude and latitude coordinates. Because tweets with GPS coordinates represent 2%-3% of all tweets and Twitter delivers tweets matching a request's parameters up to a 1% ceiling, we receive one-third to one-half all of tweets with precise location information (Morstatter et al., 2013; Leetaru, 2014).[5] We have collected these tweets in real-time, approximately five million per day, since August 26, 2013. For more information on working with Twitter data, see Steinert-Threlkeld (2018).

We then query the stored tweets to extract those from countries and days of interest. These tweets could be used for text or social network analysis, but we further select only those tweets that contain images. Twitter provides a field in each tweet called `media_url` and a flag indicating if that link is for an image. If a downloaded tweet contains an image, we retrieve it. These images form the raw material from which we generate our protest data.

We apply the protest classifier to images from periods and countries during which protest occurred. These 14 periods, shown in Table A10, generated 55,676,431 tweets containing 5,479,148 images. The classifier is applied to all 5.48 million plus the 100,000 from Google, and all images with a classification score less than .6 are dropped as they are most likely easy negatives, i.e., non-protest images. The remaining 115,060 potential protest images were then stratified based on their classification scores and sampled to ensure that the chosen images capture diverse visual features, i.e., to avoid redundant inclusion of very similar images in the dataset. This process resulted in 40,764 images that form our training set; the training set contains geolocated images from Twitter and images from Google.

3.2 Protest and Scene Classification

Step 4. Amazon Mechanical Turk provided the labor to manually annotate these 40,764 images. We asked the workers to identify the features detailed in Table A8.

Figure A1 provides examples of our AMT annotation pages. In the first task, each annotator was presented with an image and asked to judge if the image captures a protest. We assigned two

[5]For example, requesting tweets with the keyword "Microsoft" will return every tweet with that word, assuming fewer than 1% of all tweets are about Microsoft. If, however, 2% of all tweets contain that word, then Twitter will return all tweets containing that keyword until the 1% ceiling is reached. $\frac{.01}{.02} = .5$, and the same calculation is how we conclude that our corpus contains one-third to one-half of all tweets with GPS coordinates.

workers to each image and if the two workers did not agree, the image was sent to a third judge for a final verification. 11,659 of the training images contain a protest. Similarly, in the second task, annotators label the attributes listed in Table A8 that are not related to faces or violence, such as "police", "fire", "children", "flag", and so on.

As violence is a subjective and continuous variable, we used pairwise comparison annotation to generate an estimate of the *perceived violence* in an image. Among the 11,659 protest images, we randomly sampled image pairs such that each image is paired ten times. Therefore the number of pairs to be annotated was 58,295 ($11,659 \times 10 \div 2$). We then assigned ten workers for each pair and asked them to select which image looks more violent than the other. To assign the continuous violence score to each image, we use the Bradley-Terry model (Bradley and Terry, 1952) and scaled the scores to the range of $[0, 1]$. Such a pairwise comparison method usually requires more annotations but can produce more reliable and consistent ratings for subjective assessment of photographs (Kovashka et al., 2012; Joo et al., 2014, 2015). The resulting estimate for violence is therefore better conceived of as perceived violence.

Step 5. With 40,764 annotated images, we train a CNN which produces outputs for twelve variables. We used 80% of the images as the training set and the rest as the validation set. For the labels that are not face or violence related, we use a binary cross entropy (BCE) loss:

$$L_{BCE}(p, y) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log(p_n) + (1 - y_n)\log(1 - p_n)] \tag{1}$$

where p is the probability predicted by the model (CNN output for the attribute), y is the ground truth binary label (0 or 1), and N is number of images. p_n and y_n are the prediction and label for the n^{th} image, respectively.

For protester and state violence, a continuous variable, we use mean squared error (MSE) loss:

$$L_{MSE}(p, y) = -\frac{1}{N}\sum_{n=1}^{N}[(y_n - p_n)^2] \tag{2}$$

where p is the model prediction, y is the ground truth value, and N is number of images. These

are standard loss functions that are typically used in training CNNs. Note that state-violence and protester-violence are binary attributes and thus trained with a BCE loss in Eq. 1. Violence measures the degree of violence on a continuous scale, and state- and protester-violence identify the type of violence and are treated as binary variables. We use stochastic gradient descent with backpropagation to train the model. For more technical details in model training, see Won et al. (2017).

3.3 Face Classification

Step 6. We use the FairFace model developed by Kärkkäinen and Joo (2019) to classify gender, race, and age of people in images. This new model is preferred over current leading models, such as FaceNet (Schroff et al., 2015) or Face++, because it better captures race, gender, and age. Existing public face datasets and commercial APIs have been criticized for their unbalanced representation of race, as the vast majority of their face images are from people of white ethnicity (more than 80%). This results in inferior classification accuracy, especially on non-white people (Buolamwini and Gebru, 2018). Moreover, the FairFace model is trained on a large corpus of images of varying resolution, perspective, and lighting, the YFCC100M dataset (Thomee et al., 2016). This dataset is in contrast to other datasets whose images tend to be high quality, well-lit, and from the same perspective (Liu et al., 2015).

Kärkkäinen and Joo (2019) samples 102,218 of the 100 million YFCC100m images, with an explicit focus on balancing users across seven racial categories. In contrast, Liu et al. (2015) uses only three. Many other face models use skin color, but skin color is sensitive to lighting conditions. In addition, there is no other large-scale face dataset or model offers the racial category of Latino, which is critical in our study. On an external validation test, the model significantly outperforms models trained on other large-scale datasets in gender, age, and race classification.

Figure A3 shows an image from South Korea from our Twitter corpus with the face classifier applied.

4 RESEARCH DESIGN

4.1 Data

To identify protests, we searched for tweets from the fourteen periods detailed in Table A10. For each period, we searched from one week prior to the first reported protest and one week after the last one. This process identifies 55,676,431 tweets containing 5,479,148 images. To determine which to keep, we chose the lowest threshold that would maximize recall with a precision of .85. Figure A2 shows this threshold is .849 and recall is .22. This process results in 26,142 images. This number represents about one-fifth of all protest images, and 85% of them are of protest.

We then aggregate tweets to their city of origin and the day they were created. Cities are kept for analysis when at least $\frac{1}{7}$ of their days contain a protest image. Table 2 shows these 24 cities, which account for 6,303 protest images. These 6,303 protest images spread across 4,401 city days in Hong Kong, Pakistan, Spain (Catalonia only), South Korea, and Venezuela are the input for the subsequent models.[6] 1,467 of these city days contain a protest photo, so we treat the missing dates as true zeroes. A robustness check shows that this interpolation does not change results.

4.2 Bias

Using social media data frequently raises concerns about selection bias (Tufekci, 2014). If bias exists, it would come from accounts sharing images from protest activity not representative of overall protest activity. We expect that Twitter users are not representative samples of their respective countries, but we do not think the protest images they share not representative. Moreover, if the protest images are biased, structural features of the data generating process should make them less biased than any other cross-national data source. Space constraints limit us from substantiating these assertions here; see Section S2 for that substantiation.

In addition, three robustness checks presented in Section 5.2 for bias. First, we drop all tweets from “verified" accounts, which are accounts belonging to prominent individuals or organizations that Twitter has verified belong to those people or groups. Assuming they would have the most incentive to filter what they publish, removing them removes a potential source of bias. Second,

[6]The majority of images are from the United States’ Women’s March, which is not analyzed here because it does not have a dynamic component. We also recorded large numbers of tweets from Belarus and Russia, but those protests occurred on one day as well. Many more images are then at the country level, so they are discarded.

Table 2. Protest Periods

	City	Country	Start	End	Issue	Protest Images/Day	Protest Images/Day if >0
1	Central	Hong Kong	2014.09.18	2014.12.23	China reforms	1.96	5.00
2	Kowloon	Hong Kong	2014.09.18	2014.12.23	China reforms	1.29	2.92
3	Lahore	Pakistan	2017.11.07	2017.11.23	Blasphemy	.18	1
4	Kimhae	South Korea	2016.10.20	2017.03.14	Anti-incumbency	.47	1.92
5	Seoul	South Korea	2016.10.20	2017.03.14	Anti-incumbency	2.40	3.76
6	Citutat Vella	Spain	2017.09.01	2017.12.31	Secession	.94	4.95
7	Barcelona	Spain	2017.09.01	2017.12.31	Secession	3.07	11.60
8	Girona	Spain	2017.09.01	2017.12.31	Secession	1.10	3.26
9	Granera	Spain	2017.09.01	2017.12.31	Secession	.62	2.33
10	Granollers	Spain	2017.09.01	2017.12.31	Secession	.23	1.25
11	Lleida	Spain	2017.09.01	2017.12.31	Secession	.42	1.88
12	Mataro	Spain	2017.09.01	2017.12.31	Secession	.51	2.33
13	Reus	Spain	2017.09.01	2017.12.31	Secession	.35	1.68
14	Sabadell	Spain	2017.09.01	2017.12.31	Secession	.96	2.66
15	St. Cugat d. Valles	Spain	2017.09.01	2017.12.31	Secession	.31	2.06
16	St. Feliu d. Pallerols	Spain	2017.09.01	2017.12.31	Secession	.61	2.19
17	St. Salvador d. Guardiola	Spain	2017.09.01	2017.12.31	Secession	.48	2.15
18	Tarragona	Spain	2017.09.01	2017.12.31	Secession	.57	1.94
19	Terrassa	Spain	2017.09.01	2017.12.31	Secession	.57	2.22
20	Boca del Rio	Venezuela	2014.03.27	12.17.2017	Anti-Maduro	.26	1.34
21	Caracas	Venezuela	2014.03.27	12.17.2017	Anti-Maduro	4.82	7.63
22	Caucagua	Venezuela	2014.03.27	12.17.2017	Anti-Maduro	.53	1.72
23	Maracaibo	Venezuela	2014.03.27	12.17.2017	Anti-Maduro	.39	1.49
24	Valencia	Venezuela	2014.03.27	12.17.2017	Anti-Maduro	.41	1.62

we only look at tweets from accounts between the 25^{th}-75^{th} percentile of their country's follower distribution. Accounts below this range are likely to be bots or accounts which use Twitter sporadically, while accounts above this range are more likely to be strategic with their posts. Third, we remove tweets not in the *lingua franca* of their country, under the assumption that those are aimed at international audiences and so are more likely to represent a protest differently than tweets in the main language (Bruns et al., 2013).

4.3 Operationalization

The dependent variable is $Log_{10}(Protest\ Size)_{i,t}$, the logarithm of the sum of the number of faces in all protest photos from city i on day t. Other studies have found that activity on Twitter correlates with verified estimates of crowd size for airports, stadiums, and protests (Botta et al., 2015). Those estimates require either more data than were available to us or use text analysis to identify protesters. Text analysis does not scale as easily as image analysis because it requires domain expertise, so counting faces is preferred.

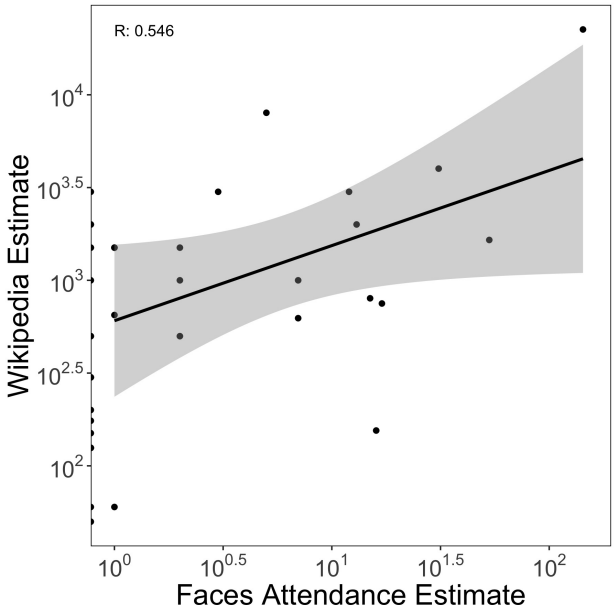
Figure 1 shows that this approach correlates with the size of protests in Russia and South Korea, as reported in newspapers (Russia) or by activists and the police (South Korea).[7] Small protests reported in other sources corresponds closely with, and without bias to the size of the protest, as what $Log_{10}(Protest\ Size)_{i,t}$ estimates. Figure 2 shows how the protest size varies over time in Catalonia, Spain and South Korea, with important events marked. There are clear spikes that correspond to major events.[8] For a verification of $Log_{10}(Protest\ Size)_{i,t}$ against protest size as recorded from cell phone location records and newspapers, see Sobolev et al. (2019). That summing the number of faces in protest photos correlates well with protests in South Korea, Russia, and the United States regardless of whether newspapers, reports from participants (activists or police), or cell phones gives us enough confidence to trust this approach in other settings.

The violence variables to test Hypothesis 1 are $Perceived\ Protester\ Violence_{i,t-1}$, $Perceived\ State\ Violence_{i,t-1}$, $Police_{i,t-1}$, and $Fire_{i,t-1}$. The violence measures are the average of the classifier

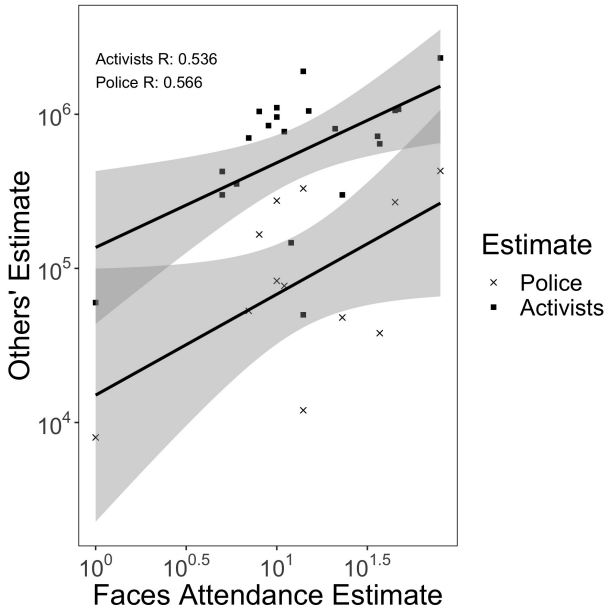
[7]Wikipedia provides the three sets of estimates.

[8]For exposition, these results are aggregated to the country-day level. A plot restricting analysis to Barcelona or Seoul shows the same trends.

Fig. 1. Verifying Dependent Variable, Cross-section

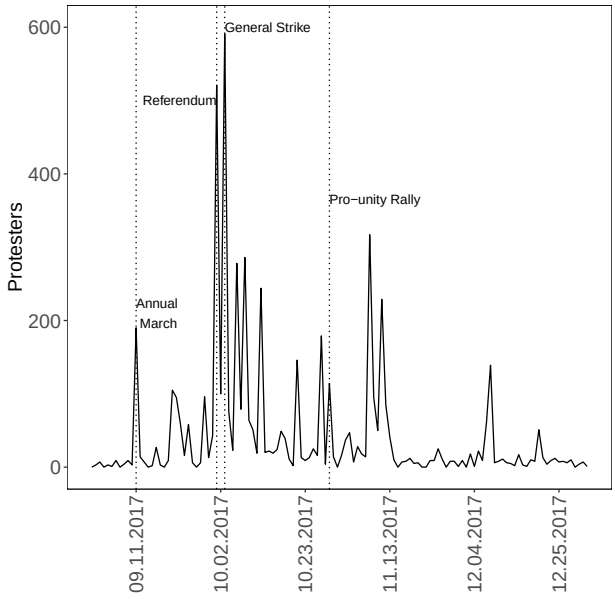


(a) Russia

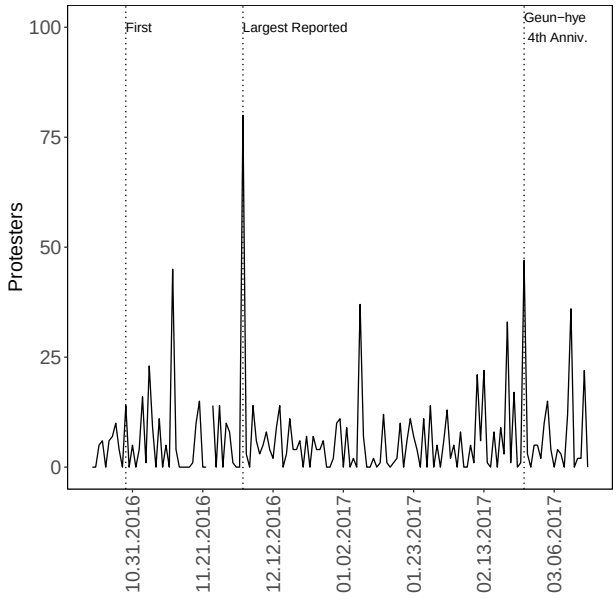


(b) South Korea

Fig. 2. Verifying Dependent Variable, Time Series



(a) Catalonia, Spain



(b) South Korea

estimate for all protest photos per city-day. The police and fire variables are the sum of images containing a police officer or fire, respectively, based on the thresholds identified in Table A9.

We describe the violence variables as “perceived" for three reasons. First, the true amount of violence is unknown because violence is not a physical entity directly measurable, like temperature or pressure. Second, the images people share may be strategically chosen. This possible selection effect is true of any event data that relies on secondary sources, which is to say almost all event data. For a longer discussion of bias that these measures may introduce, see Section 4.2. Third, the main analysis does not deduplicate images, meaning images which are shared often will have a greater impact on people’s decision making process than those only tweeted once. Deduplicating images to more closely approximate the “true" violence at events does not change results, as Table 7 shows.

Students in democracies and autocracies often spearhead mass protests (Zhao, 1998; Gonzalez, 2019). The young are more likely to lack jobs, have little wealth to lose, and view protest participation as its own end. These effects are amplified when there are many of them, a phenomenon commonly called the “youth bulge" (Urdal, 2006). Knowing that youth often make protests more intense (Hollander and Byun, 2015), states with large youth populations engage in more preventative repression (Nordås and Davenport, 2013). The percent of participants aged 20-29 is therefore a variable for which we control.

A society with greater gender equality is more likely to see nonviolent than violent action (McCammon et al., 2001; Schaftenaar, 2017), and the same is true at the movement level (Asal et al., 2013). Even when excluded from high-level leadership positions, women can play important roles as bridges between that level and the broader movement (Robnett, 1996). Women were also integral actors, as activists and participants, during the Arab Spring, a dynamic often overlooked in accounts of those events (Newsom and Lengel, 2012; Rizzo et al., 2012). The percent of protesters who are male is therefore a variable for which we control.

Figures 3 and 4 show the distribution of state violence and white faces, respectively, by country. Because most photos record no violence, Figure 3 shows the distribution after dropping all photos

whose value for *Perceived State Violence*$_{i,t-1}$ is below the median; the relative order of states does not change if all images are kept. The average amount of violence matches expectations: Venezuela's protests are frequently met with violent repression, and Spanish police aggressively met protesters. Some violence was reported in Hong Kong as the protests neared their end. No violence was reported in South Korea, and Pakistani authorities let the anti-blasphemy protests run their course. The distribution of white faces also matches expectations: Catalonia and Venezuela record the highest percentages, in that order, while Hong Kong and South Korea record almost none. The race classifier performs less well on Pakistan. Section S7 shows similar charts for protester violence and the two other demographic variables.

Fig. 3. Distribution of State Violence by Country (vertical line is the mean)

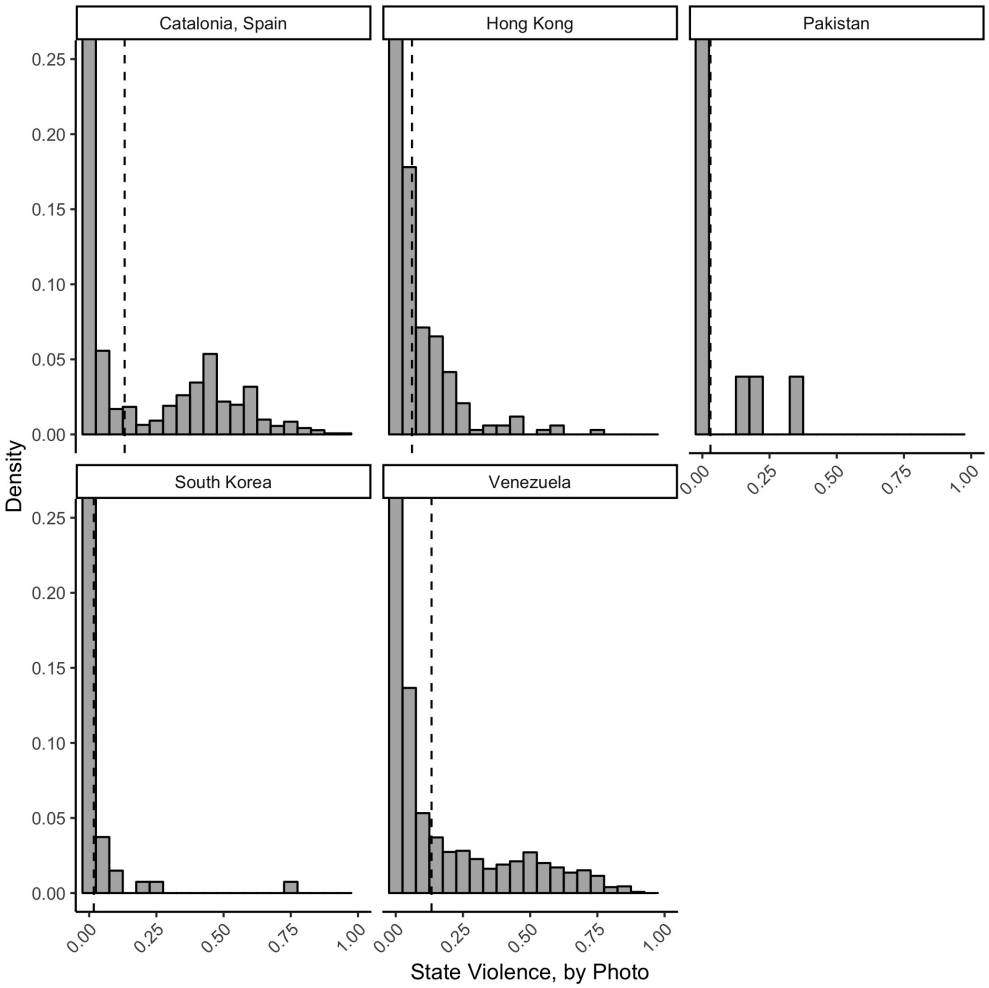


Fig. 4. Distribution of White Faces by Country (vertical line is the mean)

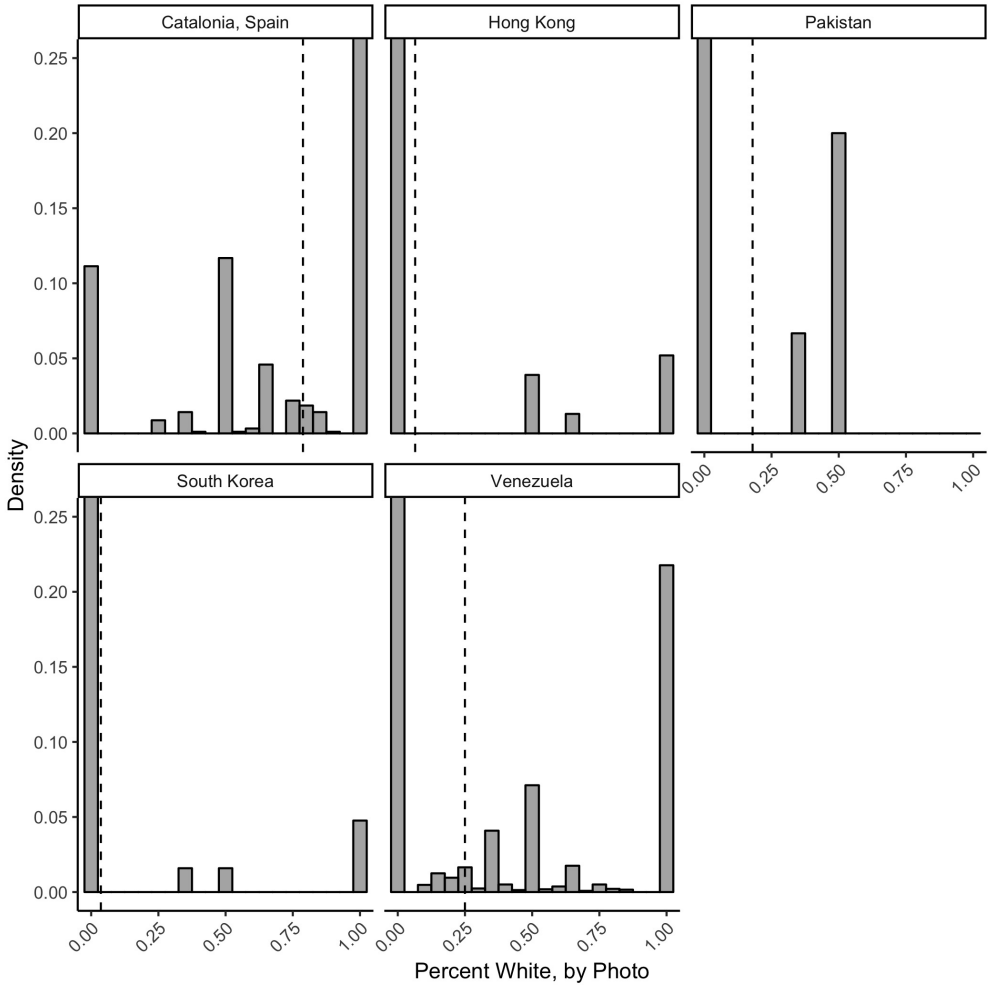


Table 3 provides descriptive statistics of these variables. Figure A9 shows the city-day correlation of these variables, and Figure A10 shows their tweet level correlation. The largest tweet level correlation, .56, is between $Race\ Diversity_{i,t-1}$ and $Age\ Diversity_{i,t-1}$. At the city-day level, the three diversity variables correlate between .7 and .78. All other correlations are below .32.

4.4 Model

In addition to the operationalizations detailed in the previous section, we include two control variables. $Tweets_{i,t-1}$ is the number of protest images per country-day and proxies for the amount of information available to protesters. This variable captures any effect general knowledge about a protest will have on protest size (Little, 2015). We also include a lagged dependent variable to

Table 3. Summary Statistics

Statistic	N	Mean	St. Dev.	Min.	Max.
$Protest\ Size_{i,t}$	4,401	2.42	14.87	0.00	627
$Perceived\ Protester\ Violence_{i,t-1}$	4,376	0.03	0.12	0.00	1.00
$Perceived\ State\ Violence_{i,t-1}$	4,376	0.02	0.07	0.00	0.94
$Police_{i,t-1}$	4,376	0.001	0.04	0.00	1.00
$Fire_{i,t-1}$	4,376	0.07	0.41	0.00	7.00
$Gender\ Diversity_{i,t-1}$	4,376	0.02	0.18	0.00	5.00
$Race\ Diversity_{i,t-1}$	4,376	0.02	0.11	0.00	0.69
$Age\ Diversity_{i,t-1}$	4,376	0.07	0.31	0.00	2.66

account for autocorrelation as well as any regression to the mean.

We build three models. The first uses only covariates that measure violence, testing H1. The second focuses on the demographic variables. The final model combine the three sets of variables. All independent variables are lagged one day. All models include city fixed effects and city-clustered standard errors, though we run robustness checks with different fixed effects and clustering.

To facilitate interpretation, ordinary least squares is the estimator. Since the dependent variable is a logarithm, the interpretation of a coefficient is the percent change in protest size as the result of a one unit increase in the independent variable. Finally, to guard against overfitting, we use five-fold cross-validation: each model is run on five different subsamples of the data and the results are averaged.

5 RESULTS

5.1 Main Results

Our models most strongly confirm the expectations for protester and state violence. Racial diversity has a signalling effect while gender diversity supports critical mass interpretations.

When protesters engage in violence, subsequent protest is smaller. Low amounts of state violence correlate with larger subsequent protests, though severe enough violence will decrease the size of protests. In addition, the more photos that show fire or police at protest, the more people mobilize. Protester violence has a much smaller slope than either state violence variable, with the

largest effect occurring when states engage in high levels of violence.

We find no statistically significant correlation between racial or gender diversity and subsequent protest size. In other models, shown soon in the robustness section and in the Supplementary Materials, gender diversity attains statistical significance with a negative slope and racial diversity does the same in the opposite direction.

Table 4. Main Result

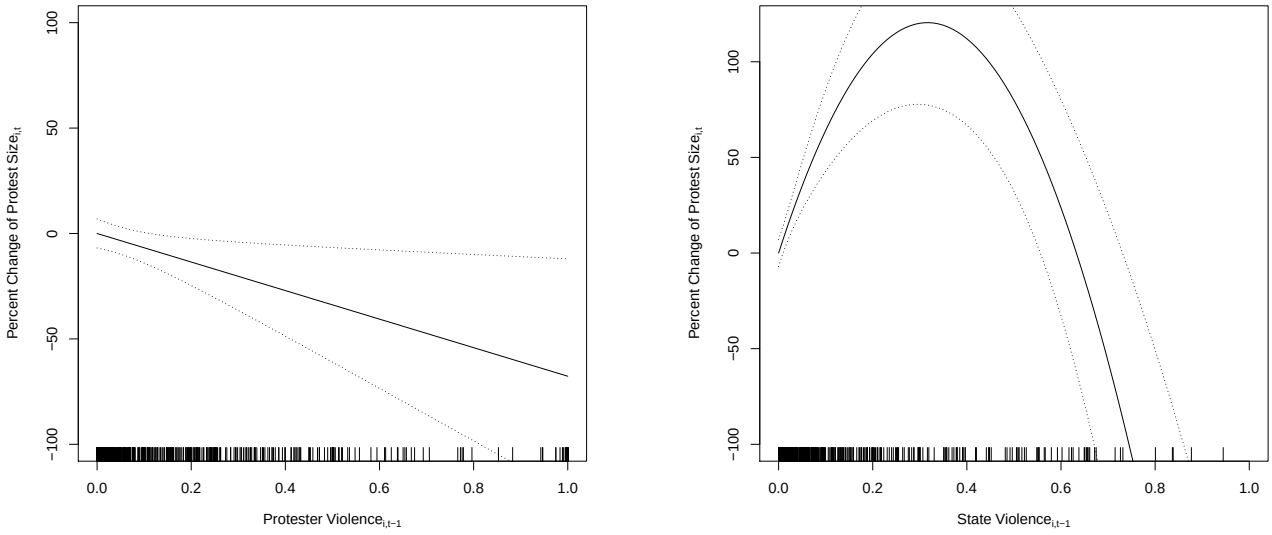
	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$		
	Violence	Demographics	Combined
	(1)	(2)	(4)
Perceived Prtstr. Violence$_{i,t-1}$	$-.1824^{**}$		$-.1674^{**}$
	(.0714)		(.0677)
Perceived Stt. Violence$_{i,t-1}$	1.2850^{**}		1.2820^{***}
	(.3152)		(.3327)
Perceived Stt. Violence$^2_{i,t-1}$	-2.0956^{***}		-2.1030^{***}
	(.5820)		(.6093)
Police$_{i,t-1}$	$.7566^{*}$		$.7626^{*}$
	(.4568)		(.4493)
Fire$_{i,t-1}$	$.1099^{***}$		$.1009^{***}$
	(.0203)		(.0236)
Gender Diversity$_{i,t-1}$		$-.1425$	$-.1126$
		(.0922)	(.0939)
Race Diversity$_{i,t-1}$		$.0955^{**}$	.0683
		(.0431)	(.0440)
Age Diversity$_{i,t-1}$		.0233	.0203
		(.0317)	(.0289)
Tweets$_{i,t-1}$	$.0093^{***}$	$.0110^{***}$	$.0095^{***}$
	(.0034)	(.0039)	(.0033)
DV$_{i,t-1}$	$.1753^{**}$	$.1828^{**}$	$.1578^{**}$
	(.0722)	(.0736)	(.0682)
Intercept	$.1227^{***}$	$.1306^{***}$	$.1260^{***}$
	(.0173)	(.0237)	(.0237)
N	4,376	4,376	4,376
City FE	Y	Y	Y
Cluster SE	Y	Y	Y
Adjusted R^2	.2435	.2280	.2450

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

City-clustered standard errors shown in parentheses.

Figure 5 shows marginal effects of protester and state violence. From values of [0-3), state violence increases protest, reaching a maximum at .3. At that amount of violence, protest size the next day is 137% higher than if there was no state violence. Moreover, state repression usually leads to larger protests: only 77 of 1,467 city-days of protest contain average state violence greater than .3. Protester violence, on the other hand, monotonically correlates with smaller protest. The change, however, is much smaller than for state violence: moving from no protester violence to its mean (.035) decreases protest size by just over 2%, while the difference between state violence and its mean is an increase of just over 17%. A one standard deviation increase of state violence from 0 increases protests size by approximately 63%; a one standard deviation increase in protester violence from the same point decreases protest size by just over 12%.

Fig. 5. Marginal Effects



(a) Protester Violence

(b) State Violence

5.2 Robustness Checks

Three sets of robustness checks confirms the results in Table 4. The first set introduces additional fixed effects and does not cluster standard errors. The second set subsets the raw data to rule out strategic behavior of Twitter users driving the results. The third set removes bots and duplicate images to confirm that results are not due to malfeasance or virality. In all presented results, the first model is the full model from Table 4.

Table 5 shows that the main findings are robust to alternate model specifications. A rule of thumb is to not cluster standard errors when there are fewer than 30, and we have 23 (Cameron et al., 2008; King and Roberts, 2015). Model 2 therefore does not cluster standard errors. Without clustering standard errors, $Gender\ Diversity_{i,t-1}$ and $Race\ Diversity_{i,t-1}$ are statistically significant. Model 3 includes a fixed effect for Saturdays and Sundays, the most popular protest days. $Gender\ Diversity_{i,t-1}$ is once again statistically significant, while $Race\ Diversity_{i,t-1}$ is not. Model 4 includes a day of week fixed effect, since some countries (primarily Venezuela and Pakistan) have larger protests outside of the weekend; results match Models 2 and 3. In case unobserved country heterogeneity drives results, we include country fixed effects. Now, $Gender\ Diversity_{i,t-1}$ just barely loses statistical significance while $Race\ Diversity_{i,t-1}$ just barely obtains it. In none of the extra checks does inference about perceived protester or state violence change.

The next set of robustness checks verify that neither strategic behavior nor data pollution drive the results. Table 6 uses different subsets of the raw data to attempt to rule out strategic behavior of individuals driving the results. The patterns hold.

One source of bias in newspaper based event data is that newspapers tailor their reporting for their intended audience. The same could be true of Twitter users, and it is more likely to be true the more likely they are to have an audience. Model 2 shows the result when keeping only tweets from accounts that fall within the 25^{th}-75^{th} percentile of their country's follower distribution. The coefficient on $Gender\ Diversity_{i,t-1}$ remains the same, but it is not statistically significant because of a smaller standard error. Other results are not susceptible to keeping users based on their follower count. Model 3 restricts the sample to only users who Twitter has not verified; for

Table 5. Robust to Alternate Specifications

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$				
	Original Model	Original Model	Weekend FE	Day-of-week FE	Country FE
	(1)	(2)	(3)	(4)	(5)
Perceived Prtstr. Violence$_{i,t-1}$	−.1674**	−.1674***	−.1675***	−.1693**	−.1579**
	(.0677)	(.0555)	(.0806)	(.0807)	(.0793)
Perceived Stt. Violence$_{i,t-1}$	1.2820***	1.2820***	1.2857***	1.3001***	1.3824***
	(.3327)	(.2158)	(.3668)	(.3673)	(.3654)
Perceived Stt. Violence$^2_{i,t-1}$	−2.1030***	−2.1030***	−2.1042***	−2.1241***	−2.2609***
	(.6093)	(.3583)	(.6493)	(.6504)	(.6353)
Police$_{i,t-1}$	.7626*	.7626***	.7688***	.7729*	.7708*
	(.4493)	(.1489)	(.4627)	(.4613)	(.4436)
Fire$_{i,t-1}$	.1009***	.1009***	.1012***	.1014***	.1093***
	(.0236)	(.0161)	(.0375)	(.0376)	(.0384)
Gender Diversity$_{i,t-1}$	−.1126	−.1126***	−.1152***	−.1146*	−.1023
	(.0939)	(.0415)	(.0638)	(.0636)	(.0650)
Race Diversity$_{i,t-1}$	.0683	.0683**	.0712	.0718	.0912*
	(.0440)	(.0280)	(.0463)	(.0464)	(.0468)
Age Diversity$_{i,t-1}$	.0203	.0203	.0213	.0206	.0127
	(.0289)	(.0230)	(.0350)	(.0349)	(.0357)
Tweets$_{i,t-1}$	.0095***	.0095***	.0095***	.0094***	.0093***
	(.0033)	(.0009)	(.0033)	(.0033)	(.0033)
DV$_{i,t-1}$	.1578***	.1578***	.1559***	.1564***	.1901***
	(.0682)	(.0237)	(.0405)	(.0405)	(.0408)
Intercept	.1260***	.1260***	.1186***	.1277***	.0885***
	(.0237)	(.0158)	(.0186)	(.0216)	(.0088)
N	4,376	4,376	4,376	4,376	4,376
Adjusted R^2	.2450	.2450	.2459	.2459	.2240
City FE	Y	Y	Y	Y	N
Weekend FE	N	N	Y	N	N
Weekday FE	N	N	N	Y	N
Clustered SE	Y	N	Y	Y	Y

*p < .1; **p < .05; ***p < .01

City-clustered standard errors shown in parentheses for Models 1, 3, and 4. Model 5 uses country-clustered standard errors.

public figures such as celebrities or politicians (and even some academics), Twitter verifies that the account actually belongs to the person it purports to. Since these users should be more likely to engage with Twitter strategically, we drop their tweets from analysis. This process takes away 447 tweets and 5 city-days, which is enough to reduce the coefficient of Police$_{i,t-1}$ below traditional thresholds of statistical significance; all other results match the full model.

The results when keeping tweets only in a country's *lingua franca*, shown in Model 4 of Table 6, are particularly interesting. Many Twitter users change their language depending on political context (Metzger et al., 2015), often as a method of attracting foreign audiences (Bruns et al.,

2013). For the violence and demographic variables except $Race\ Diversity_{i,t-1}$, the coefficients are much larger than the original model. In addition to violence now being estimated to have a stronger effect, $Gender\ Diversity_{i,t-1}$ is 80% larger and its standard error is halved, making it statistically significant. $Race\ Diversity_{i,t-1}$'s coefficient decreases, but its standard error decreases by even more, making it statistically significant as well. The most noticeable change is to $Age\ Diversity_{i,t-1}$, whose point estimate triples while its standard error remains the same. In only two other models, shown in the Supplementary Materials, is this variable statistically significant. Overall, restricting by language produces a model with a 22.5% better fit than the original. This better fit, larger coefficients, and more precise estimates of those coefficients suggests that language use may be one of the most common ways users behave strategically on social media.

Table 6. Robust to Strategic Behavior

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$			
	Original	Normal Users	No Verified Accounts	Country's Language
	(1)	(2)	(3)	(4)
Perceived Prtstr. Violence$_{i,t-1}$	−.1674***	−.1616***	−.1438**	−.1899***
	(.0677)	(.0422)	(.0670)	(.0425)
Perceived Stt. Violence$_{i,t-1}$	1.2820***	1.2564***	1.3146***	1.4723***
	(.3327)	(.3347)	(.3597)	(.4686)
Perceived Stt. Violence$^2_{i,t-1}$	−2.1030***	−2.1084***	−2.0819***	−2.4132***
	(.6093)	(.6093)	(.6156)	(.9056)
Police$_{i,t-1}$	.7626*	.8606**	.6409	.4802
	(.4493)	(.3737)	(.3930)	(.2929)
Fire$_{i,t-1}$	.1009***	.0613*	.0876***	.0664***
	(.0236)	(.0331)	(.0234)	(.0252)
Gender Diversity$_{i,t-1}$	−.1126	−.1124*	−.1121	−.1820***
	(.0939)	(.0662)	(.0901)	(.0592)
Race Diversity$_{i,t-1}$	.0683	.0316	.0675	.0504*
	(.0440)	(.0339)	(.0411)	(.0295)
Age Diversity$_{i,t-1}$	.0203	.0083	.0149	.0773***
	(.0289)	(.0293)	(.0285)	(.0282)
Tweets$_{i,t-1}$	.0095***	.0156***	.0124**	.0259***
	(.0033)	(.0060)	(.0048)	(.0044)
DV$_{i,t-1}$	.1578***	.1221**	.1412**	.0591
	(.0682)	(.0562)	(.0685)	(.0706)
Intercept	.1260***	.1283***	.1185***	.0701***
	(.0237)	(.0156)	(.0155)	(.0137)
N	4,376	3,715	4,371	3,614
Adjusted R^2	.2450	.1759	.2457	.3002
City FE	Y	Y	Y	Y
Cluster SE	Y	Y	Y	Y

*p < .1; **p < .05; ***p < .01

Table 7 presents two attempts to ensure that results are not driven by quirks in the data generation process. The prevalence of bots - social media accounts controlled by computer code - has raised concerns about the veracity of studies relying on social media data (Ferrara and Bessi, 2016). Though other work has found few bots in geolocated tweets (Driscoll and Steinert-Threlkeld, 2018), we nonetheless submit every user to the Botometer service and remove tweets with a complete automation probability $\geq .4$, the threshold which has been found to produce the most accurate classification of bots (Varol et al., 2017). Model 2 presents these results, and findings do not change. See Table A11 in the Supplementary Materials for the percent of accounts and tweets that are from bots, by country; no more than 6.5% of tweets in any country are from bots.

To confirm that repetition of images do not drive results, we remove duplicate images. Models 3 and 4 from Table 7 shows these results. While our data do not contain retweets because Twitter does not assign coordinates to retweets, they do contain replies, and replies contain the image of the original tweet. (Section S8.2 details this methodology, and Table A12 show the percent of tweets per city that are duplicates.) This process removes 2,920 images from the periods in question, and Model 3 presents the results. Results for the violence and demographic variables do not change. Model 4 weights these data by the number of protest tweets per city-day. In this model, the coefficients for the the perceived violence variable are up to twice as large as the original model, though the coefficients for the demographic variables shrink.[9] Model 4 also produces the best fit of any model we build. The violence and demographic conclusions are not affected by bots or the reproduction of images: our model's original measurements do not appear to measure perception as much as they do actual effects.

The Supplementary Materials present seven additional sets of robustness checks in Tables A13 through A19. The first set changes the operationalization of the dependent variable. The second uses 15 lags of the dependent variable, as suggested by a partial autocorrelation plot. The third set uses count models, and the fourth weights city-days by their number of tweets. The fifth set increases the probability that tweets are from a protest by discarding those not from mobile devices

[9]Note as well that Race Diversity$_{i,t-1}$ becomes negative, but with a small p-value, in the deduplicated models.

Table 7. Robust to Pollution

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$			
	Original	No Bots	Deduplicate Images	Deduplcate Images, Weighted
	(1)	(2)	(3)	(4)
Perceived Prtstr. Violence$_{i,t-1}$	−.1674***	−.1593**	−.1659*	−.3778**
	(.0677)	(.0649)	(.0921)	(.1513)
Perceived Stt. Violence$_{i,t-1}$	1.2820***	1.2461***	1.2613**	2.3127***
	(.3327)	(.3326)	(.4913)	(.4773)
Perceived Stt. Violence$^2_{i,t-1}$	−2.1030***	−2.0535***	−2.0571***	−4.1393***
	(.6093)	(.5879)	(.8778)	(.8678)
Police$_{i,t-1}$	.7626*	.6582*	.9543*	1.4324***
	(.4493)	(.3806)	(.5359)	(.2220)
Fire$_{i,t-1}$	.1009***	.0923***	.0735	.0178
	(.0236)	(.0268)	(.0634)	(.0210)
Gender Diversity$_{i,t-1}$	−.1126	−.1231	.0023	−.0199
	(.0939)	(.0862)	(.0411)	(.0618)
Race Diversity$_{i,t-1}$	.0683	.0623	−.0115	−.0197
	(.0440)	(.0390)	(.0434)	(.0393)
Age Diversity$_{i,t-1}$	.0203	.0238	.0215	.0046
	(.0289)	(.0301)	(.0240)	(.0341)
Tweets$_{i,t-1}$	.0095***	.0127***	.0173**	.0067***
	(.0033)	(.0048)	(.0073)	(.0008)
DV$_{i,t-1}$	.1578***	.1396**	.1221	.1471***
	(.0682)	(.0641)	(.0874)	(.0482)
Intercept	.1260***	.1214	.1328	.4891***
	(.0237)	(.0156)	(.0169)	(.0586)
N	4,376	4,376	2,786	2,786
Adjusted R^2	.2450	.2508	.1912	.4479
City FE	Y	Y	Y	Y
Cluster SE	Y	Y	Y	Y

*p < .1; **p < .05; ***p < .01

or from non-protest hours. The sixth runs the full model separately by country, and the seventh investigates how Police$_{i,t-1}$ and Fire$_{i,t-1}$ correlate with the perceived violence measures.

6 DISCUSSION

6.1 Images and Measurement

Emphasizing the severity of repression during protest policing is not new (Muller, 1985; Khawaja, 1993); measuring repression as a continuous variable is. For example, the Social Conflict Analysis Database (SCAD), Urban Social Disorder, and Armed Conflict Locations and Event Data (ACLED) dataset record repression during an event as occurring or not (Raleigh et al., 2010;

Salehyan et al., 2012; Urdal and Hoelscher, 2012). Repression is sometimes coded as ordinal or nominal as well (Goldstein, 1992; Stephan and Chenoweth, 2008; Clark and Regan, 2016), and machine-coded event data like the Integrated Conflict Early Warning System use this approach (Gerner et al., 2002; Boschee et al., 2015).

As far as we are aware, all previous approaches generate nominal or ordinal repression variables from primary or secondary sources. This process, completely understandable given how violence is recorded in texts, creates an implicit mapping of a latent quantity onto discrete categories. This mapping is problematic because each researcher has its own mental model, so different studies are likely to map the same latent quantity onto different discrete categories (values of the ordinal variable). Measuring repression on a continuous scale may therefore provide a clearer understanding of how it affects protest dynamics. It also facilitates the inclusion and interpretation of interaction terms for violence, allowing us to test for nonlinear effects (Moore, 1998; Shellman et al., 2013).

The results presented here suggest that measuring violence as a continuous variable may help resolve the repression-dissent puzzle. Mapping violence into discrete bins may be especially pernicious with panel data, explaining why those studies tend to find no correlation between repression and protest. We avoid this pitfall by presenting human coders with over 10,000 pairs of images to label and training a deep learning computer vision model on this training set; the model outputs continuous estimates of protester and state violence, mitigating concerns that a result for repression or protester violence is due to researcher effects. The results presented here are continuous measurements based on primary sources.

Using images generated from social media also allows for more precise temporal measurement. A difficulty testing protest dynamics is that action occurs on a timescale difficult to measure with newspaper reports, the primary source of data for these types of studies (Earl et al., 2004). Most research has therefore analyzed protest dynamics with coarse time scales such as weekly (Lohmann, 1994a; Rasler, 1996) or, usually in the case of surveys, without a time component (Opp and Gern, 1993; Beissinger, 2013). Recent research takes advantage of new datasets, including social media data, to measure protest dynamics at a daily level (Larson et al., 2016; Ritter and Conrad, 2016;

Hsuan et al., 2017; Steinert-Threlkeld, 2017). Combining this high level of resolution with the additional information that can be extracted from images has only been attempted twice before (Won et al., 2017; Zhang and Pan, 2019), as far as we are aware, though there is work at scale analyzing how the emotional content of images affects online mobilization (Casas and Webb Williams, 2018) or how different news outlets portray Black Lives Matter protests (Torres, 2018).[10]

6.2 Ethics

The advances in scholarly understanding that the combination of computer vision and social media enable also raises serious ethical concerns. We briefly discuss some and point the reader to (Joo and Steinert-Threlkeld, 2018) for a longer analysis.

Like any measurement, the results are only as good as the input data. Many off the shelf computer vision programs reproduce racial biases, and the leading datasets used to train race classifiers have relatively small corpuses of images (Grush, 2015; Lam et al., 2018). The model we use, FairFace, is less biased than other ones, however, because its training data were constructed on racially balanced images whose quality more closely resembles social media photographs than previous datasets'.

Treating race as a distinct category around which people may organize is itself problematic. We simply note that in many countries, race encapsulates a multitude of historic power imbalances. While we do not mean to reify race, ignoring it would also do a disservice to its importance in many countries' politics.

Protesters may not be as anonymous as they think. Though these data are observational and publicly available, individuals in photographs may not have consented to appear in those photographs. While true of any images of public spaces, the concern is heightened when individuals are engaged in risky behavior. Authorities could monitor images shared on social media to identify people who protested, much as some do with cell phone location data (Davenport, 2014).[11] Foreign governments and parts of the United States law enforcement already monitor faces in crowds

[10] See Cowart et al. (2016) for an example of manual protest image analysis.

[11] The logic works in reverse. Shared protest images can be used to identify incriminating state behavior that would otherwise be denied (Lim, 2013).

(Purdy, 2018; Shaban, 2018). This concern about facial recognition also means that individuals who appear in photos but who did not take the photo may not realize they can be implicated in protest.[12] To prevent the identification of individuals in our data, we have chosen not to release the tweet identification number or image URL for the raw data.

6.3 Lower Bounds

This paper's findings are appealing because the precision and resolution of its measures allows for deeper theoretical understanding of protest dynamics. Moreover, the precision of these results should be considered a lower bound, as the number of protest images located to the city level is quite small. Researchers can increase the number of images available, and therefore analyze more events with more precision, using five tactics.

First, the easiest approach would be to accept a less rigorously defined measure of location, users' self-reported location. Twitter profiles contain a location field that individuals can populate with any phrase, e.g. "Los Angeles, CA" or "A Server Somewhere". Approximately 75% of Twitter users have information in this field, but only 8% of users (in the United States) have a string that Google can resolve to a specific location (Mislove et al., 2011). (Globally, $\frac{1}{3}$ of accounts have location or profile information in English (Leetaru et al., 2013).) In the United States, approximately 3.4% of accounts enable GPS coordinates, so using the location field at least doubles the number of images available (Sloan and Morgan, 2015). This increase will be more pronounced the less frequently a country's users geotag: only .3% and .9% of tweets in Korean and Arabic, respectively, contain GPS coordinates (Sloan and Morgan, 2015).

Second, one could purchase tweets from a vendor or download past tweets of users who tweet from protest events.

Third, newspapers and television channels maintain Twitter accounts and share the same stories there that are used in event datasets. While their accounts are not, in our experience, geolocated, it is theoretically feasible to incorporate the articles they share into an event data generating pipeline. Doing so would allow the researcher to determine the sensitivity of event records to source type,

[12]The UCLA IRB approved this study. We emphasize again that we only use publicly available data.

more precisely measure bias from sources, and determine if events recorded from traditional media have different effects than those from social media.

Two other approaches would move beyond Twitter to collect images. The fourth approach should look to other online platforms, especially Instagram, to collect images. Instagram provides much less data through its application programming interface than Twitter does, so one will have to crawl it. Crawling is a more technically difficult procedure and is actively discouraged. Instagram tends to be used for apolitical postings as well. Flickr has been used to track protests, but it is not a widely used platform (Alanyali et al., 2016). The fifth approach would be to partner with a news images provider, such as The Associated Press or Getty Images.

6.4 Conclusion

We have presented results on a relatively small number of protests, and future work should increase the number of protests analyzed. Doing so will rely on luck, as protests will have to occur in countries use Twitter, or other platforms, heavily. Developing infrastructure to collect more tweets with images will decrease the role luck plays in directing research.

Because images contain more information than text, they hold much promise for the study of phenomena of interest to political scientists. For generating event data, images hold particular promise in measuring magnitude, both in terms of crowd size and severity of an event, as well as reducing bias from newspaper data (Sobolev et al., 2019). For explanations of how these data can contribute to subfields like political behavior, communication, or international relations, see Joo and Steinert-Threlkeld (2018). The techniques for generating useful data are similar and different to text-as-data approaches, and this paper demonstrates one area in which computer vision techniques benefit political scientists. Future work should expand on the data and variables introduced in this paper.

REFERENCES

Alanyali, M., Preis, T., and Moat, H. S. (2016). "Tracking Protests Using Geotagged Flickr Photographs." *PLoS ONE*, 11(3), 27–30.

Asal, V., Legault, R., Szekely, O., and Wilkenfeld, J. (2013). "Gender ideologies and forms of contentious mobilization in the Middle East." *Journal of Peace Research*, 50(3), 305–318.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "Openface: an open source facial behavior analysis toolkit." *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, IEEE, 1–10.

Bashir, N. Y., Lockwood, P., Chasteen, A. L., Nadolny, D., and Noyes, I. (2013). "The ironic impact of activists: Negative stereotypes reduce social change influence." *European Journal of Social Psychology*, 43(7), 614–626.

Battaglini, M. (2017). "Public Protests and Policy Making." *Quarterly Journal of Economics*, 132(1), 485–549.

Baum, M. A. and Zhukov, Y. M. (2015). "Filtering revolution: Reporting bias in international newspaper coverage of the Libyan civil war." *Journal of Peace Research*, 52(3), 384–400.

Baum, M. A. and Zhukov, Y. M. (2018). "Media Ownership and News Coverage of International Conflict1." *Political Communication*, 1–28.

Beissinger, M. R. (2013). "The Semblance of Democratic Revolution: Coalitions in Ukraine's Orange Revolution." *American Political Science Review*, 107(03), 574–592.

Benford, R. D. and Snow, D. A. (2000). "Framing Processes and Social Movements: An Overview and Assessment." *Annual Review of Sociology*, 26, 611–639.

Bergamo, A. and Torresani, L. (2010). "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach." *Advances in neural information processing systems*, 181–189.

Biggs, M. (2003). "Positive feedback in collective mobilization: The American strike wave of 1886." *Theory and Society*, 32, 217–254.

Biggs, M. (2016). "Size Matters: Quantifying Protest by Counting Participants." *Sociological Methods & Research*, 1–33.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). "End to end learning for self-driving cars." *arXiv preprint arXiv:1604.07316*.

Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., and Ward, M. (2015). "ICEWS Coded Event Data, <http://dx.doi.org/10.7910/DVN/28075>.

Botta, F., Moat, H. S., and Preis, T. (2015). "Quantifying crowd size with mobile phone and Twitter data." *Royal Society Open Science*, 2, 150162.

Bradley, R. A. and Terry, M. E. (1952). "Rank analysis of incomplete block designs: I. the method of paired comparisons." *Biometrika*, 39(3/4), 324–345.

Bratton, M. and Walle, N. V. D. (1992). "Popular Protest and Political Reform in Africa." *Comparative Politics*, 24(4), 419–442.

Bruns, A., Highfield, T., and Burgess, J. (2013). "The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks." *American Behavioral Scientist*, 57(7), 871–898.

Bueno de Mesquita, B., Smith, A., Siverson, R. M., and Morrow, J. D. (2003). *The Logic of Political Survival*. MIT Press, Cambridge.

Buolamwini, J. and Gebru, T. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on Fairness, Accountability and Transparency*, 77–91.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3), 414–427.

Cantu, F. "The Fingerprints of Fraud: Evidence From Mexico's 1988 Presidential Election." *American Political Science Review*, Forthcoming.

Carey, S. C. (2006). "The Dynamic Relationship Between Protest and Repression." *Political Research Quarterly*, 59(1), 1–11.

Casas, A. and Webb Williams, N. (2018). "Images That Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly*.

Celestino, M. R. and Gleditsch, K. S. (2013). "Fresh carnations or all thorn, no rose? Nonviolent campaigns and transitions in autocracies." *Journal of Peace Research*, 50(3), 385–400.

Chenoweth, E. and Schock, K. (2015). "Do Contemporaneous Armed Challenges Affect the Outcomes of Mass Nonviolent Campaigns?." *Mobilization: An International Quarterly*, 20(4), 427–451.

Chenoweth, E. and Stephan, M. J. (2011). *Why Civil Resistance Works*. Columbia University Press, New York City.

Clark, D. H. and Regan, P. M. (2016). "Mass Mobilization, <https://www.binghamton.edu/massmobilization/about.html>.

Cowart, H. S., Saunders, L. M., and Blackstone, G. E. (2016). "Picture a Protest: Analyzing Media Images Tweeted From Ferguson." *Social Media and Society*, 2(4), 1–9.

Danneman, N. and Ritter, E. H. (2013). "Contagious Rebellion and Preemptive Repression." *Journal of Conflict Resolution*, 58(2), 254–279.

Davenport, C. (2007). "State Repression and Political Order." *Annual Review of Political Science*, 10(1), 1–23.

Davenport, C. (2014). "Old Wine in an E-bottle (or, the Text that Mistook Itself for a Tactical Shift), <http://politicalviolenceataglance.org/2014/01/28/old-wine-in-an-e-bottle-or-the-text-that-mistook-itself-for-a-tactical-shift/>.

Davenport, C. and Armstrong II, D. A. (2004). "Democracy and the Violation of Human Rights: A Statistical Analysis from 1976 to 1996." *American Journal of Political Science*, 48(3), 538–554.

Davenport, C. and Soule, S. A. (2009). "Velvet Glove, Iron Fist or Even Hand? Protest Policing in the United States, 1960-1990." *Mobilization*, 14(1), 1–22.

D. Della Porta and H. R. Reiter, eds. (1998). *Policing protest: The control of mass demonstrations in Western democracies*. University of Minnesota Press, Minneapolis.

DeNardo, J. (1985). *Power in Numbers: The Political Strategy of Protest and Rebellion*. Princeton University Press, Princeton.

Deng, J., Dong, W., Socher, R., Li, L.-j., Li, K., and Fei-fei, L. (2009). "ImageNet : A Large-Scale Hierarchical Image Database." *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Downs, A. (1957). *An Economic Theory of Democracy*. Harper and Row, New York City.

Driscoll, J. and Steinert-Threlkeld, Z. C. (2018). "Does Social Media Enable Irredentist Information Warfare?." *Working paper*.

Earl, J. (2003). "Tanks, Tear Gas, and Taxes : Toward a Theory of Movement Repression." *Sociological Theory*, 21(1), 44–68.

Earl, J., Martin, A., Mccarthy, J. D., and Soule, S. A. (2004). "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology*, 30, 65–80.

Earl, J., McKee Hurwitz, H., Mejia Mesinas, A., Tolan, M., and Arlotti, A. (2013). "This Protest Will Be Tweeted: Twitter and protest policing during the Pittsburgh G20." *Information, Communication & Society*, 16(4), 459–478.

Fassiotto, M. and Soule, S. A. (2017). "Loud and Clear: the Effect of Protest Signals on Congressional Attention." *Mobilization: An International Quarterly*, 22(1), 17–38.

Feinberga, M., Willer, R., and Kovacheff, C. (2017). "Extreme Protest Tactics Reduce Popular Support for Social Movements.

Ferrara, E. and Bessi, A. (2016). "Social bots distort the 2016 U.S. Presidential election online discusion." *First Monday*, 21(11), 1–17.

Fisher, D. R., Dow, D. M., and Ray, R. (2017). "Intersectionality takes it to the streets : Mobilizing across diverse interests for the Women' s March." *Science Advances*, 3, 1–8.

Francisco, R. A. (1995). "The Relationship between Coercion and Protest: An Empirical Evaluation in Three Coercive States." *Journal of Conflict Resolution*, 39(2), 263–282.

Francisco, R. A. (1996). "Coercion and Protest: An Empirical Test in Two Democratic States." *American Journal of Political Science*, 40(4), 1179–1204.

Francisco, R. A. (2004). "After the Massacre: Mobilization in the Wake of Harsh Repression." *Mobilization: An International Journal*, 9(2), 107–126.

Gause, L. (2018). "The Advantage of Disadvantage: Legislative Responsiveness to Collective Action by the Politically Marginalized.

Gerner, D. J., Schrodt, P. A., Abu-Jabr, R., and Yilmaz, O. (2002). "Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions." *Annual Meeting of the International Studies Association*.

Girshick, R. (2015). "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Goldstein, J. S. (1992). "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution*, 36(2), 369–385.

Goldstone, J. A. (2001). "Toward a Fourth Generation of Revolutionary Theory." *Annual Review of Political Science*, 4, 139–187.

Gonzalez, F. (2019). "Collective action in networks: Evidence from the Chilean student movement." *Working paper*.

Graber, D. A. (1996). "Say It with Pictures." *The ANNALS of the American Academy of Political and Social Science*, 546(1), 85–96.

Grush, L. (2015). "Google engineer apologizes after Photos app tags two black people as gorillas." *The Verge*, <https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas>.

Güler, R. A., Neverova, N., and Kokkinos, I. (2018). "Densepose: Dense human pose estimation in the wild." *arXiv preprint arXiv:1802.00434*.

Gunitsky, S. (2015). "Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability." *Perspectives on Politics*, 13(01), 42–54.

Gupta, D. K., Singh, H., and Sprague, T. (1993). "Government Coercion of Dissidents: Deterrence or Provocation?." *Journal of Conflict Resolution*, 37(2), 301–339.

Gurr, T. (1970). *Why Men Rebel*. Princeton University Press, Princeton.

Gurr, T. R. and Moore, W. H. (1997). "Ethnopolitical Rebellion: A Cross-Sectional Analysis of the 1980s with Risk Assessments for the 1990s." *American Journal of Political Science*, 41(4), 1079–1103.

He, K., Zhang, X., Ren, S., and Shun, J. (2016a). "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heaney, M. T. and Rojas, F. (2008). "Coalition Dissolution, Mobilization, and Network Dynamics in the U.S. Antiwar Movement." *Research in Social Movements, Conflicts, and Change*, 28(08), 39–82.

Hellmeier, S., Weidmann, N. B., and Geelmuyden Rød, E. (2018). "In The Spotlight:Analyzing Sequential Attention Effects in Protest Reporting." *Political Communication*, 00(00), 1–25.

Hess, D. and Martin, B. (2006). "Repression, Backfire, and the Theory of Transformative Events." *Mobilization: An International Journal*, 11(2), 249–267.

Hollander, E. J. and Byun, C. C. (2015). "Explaining the Intensity of the Arab Spring." *Digest of Middle East Studies*, 24(1), 26–46.

Hsuan, T., Chen, Y., Zachary, P., and Fariss, C. J. (2017). "Who Protests? Using Social Media Data to Estimate How Social Context Affects Political Behavior.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., et al. (2015). "An empirical evaluation of deep learning on highway driving." *arXiv preprint arXiv:1504.01716*.

Joo, J., Li, W., Steen, F. F., and Zhu, S.-C. (2014). "Visual persuasion: Inferring communicative intents of images." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 216–223.

Joo, J., Steen, F. F., and Zhu, S.-C. (2015). "Automated facial trait judgment and election outcome prediction: Social dimensions of face." *Proceedings of the IEEE international conference on computer vision*, 3712–3720.

Joo, J. and Steinert-Threlkeld, Z. C. (2018). "Image as Data: Automated Visual Content Analysis for Political Science." *Working paper*.

Kalyvas, S. N. (2004). "The Urban Bias in Research on Civil Wars." *Security Studies*, 13(3), 160–190.

Kärkkäinen, K. and Joo, J. (2019). "Fairface: Face attribute dataset for balanced race, gender, and age." *arXiv preprint arXiv:1908.04913*.

Kern, H. L. (2011). "Foreign Media and Protest Diffusion in Authoritarian Regimes: The Case of the 1989 East German Revolution." *Comparative Political Studies*, 44(9), 1179–1205.

Khawaja, M. (1993). "Repression and Popular Collective Action: Evidence from the West Bank." *Sociological Forum*, 8(1), 47–71.

King, G. and Roberts, M. E. (2015). "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis*, 23(2), 159–179.

Koopmans, R. (1993). "The Dynamics of Protest Waves: West Germany, 1965 to 1989." *American Sociological Review*, 58(5), 637–658.

Kovashka, A., Parikh, D., and Grauman, K. (2012). "Whittlesearch: Image search with relative attribute feedback." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2973–2980.

Kuran, T. (1989). "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice*, 61(1), 41–74.

Lam, O., Wojcik, S., Broderick, B., and Hughes, A. (2018). "Gender and Jobs in Online Image Searches." *Report no.*, Pew Research Center.

Larson, J. M., Nagler, J., Ronen, J., and Tucker, J. A. (2016). "Social Networks and Protest Participation: Evidence from 130 Million Twitter Users.

Lawrence, A. K. (2016). "Repression and Activism among the Arab Spring's First Movers: Evidence from Morocco's February 20th Movement." *British Journal of Political Science*, (May), 1–20.

Leetaru, K. H. (2014). "Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia." *D-Lib Magazine*, 18(9), 1–16.

Leetaru, K. H., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). "Mapping the global Twitter heartbeat: The geography of Twitter." *First Monday*, 18(5-6), 1–33.

Lim, M. (2013). "Framing Bouazizi: 'White lies', hybrid network, and collective/connective action in the 2010-11 Tunisian uprising." *Journalism*, 14(7), 921–941.

Little, A. T. (2015). "Communication Technology and Protest." *Journal of Politics*, 78(1), 152–166.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep learning face attributes in the wild." *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Lohmann, S. (1993). "A Signaling Model of Informative and Manipulative Political Action." *American Political Science Review*, 87(2), 319–333.

Lohmann, S. (1994a). "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany 1989-91." *World Politics*, 47(1).

Lohmann, S. (1994b). "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-91." *World Politics*, 47(1), 42–101.

Malik, M. M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). "Population Bias in Geotagged Tweets." *9th International AAAI Conference on Weblogs and Social Media*, 18–27.

McCammon, H. J., Campbell, K. E., Granberg, E. M., and Mowery, C. (2001). "How Movements Win: Gendered Opportunity Structures and U.S. Women's Suffrage." *American Sociological Review*, 66(1), 49–70.

Mccarthy, J. D., McPhail, C., and Smith, J. (1996). "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations." *American Sociological Review*, 61(3), 478–499.

Mellon, J. and Prosser, C. (2017). "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users." *Research & Politics*, 4(3), 205316801772000.

Metzger, M., Nagler, J., and Tucker, J. a. (2015). "Tweeting Identity? Ukrainian, Russian, and #Euromaidan." *Journal of Comparative Economics*, 44(1), 16–40.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). "Understanding the Demographics of Twitter Users." *Proceedings of the Fifth International AAI Conference on the Weblogs and Social Media*, 554–557.

Moore, W. H. (1995). "Rational Rebels: Overcoming the Free-Rider Problem." *Political Research Quarterly*, 48(2), 417–454.

Moore, W. H. (1998). "Repression and Dissent: Substitution, Context, and Timing." *American Journal of Political Science*, 42(3), 851–873.

Moore, W. H. (2000). "The Repression of Dissent: A Substitution Model of Government Coercion." *Journal of Conflict Resolution*, 44(1), 107–127.

Morstatter, F., Pfeffer, J., Carley, K. M., and Liu, H. (2013). "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." *Association for the Advancement of Artificial Intelligence*.

Muller, E. N. (1985). "Income Inequality, Regime Repressiveness, and Political Violence." *American Sociological Review*, 50(1), 47–61.

Murdie, A. and Purser, C. (2017). "How protest affects opinions of peaceful demonstration and expression rights." *Journal of Human Rights*, 16(3), 351–369.

Myers, D. J. and Caniglia, B. S. (2004). "All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969." *American Sociological Review*, 69, 519–543.

Newsom, V. A. and Lengel, L. (2012). "Arab Women, Social Media, and the Arab Spring: Applying the framework of digital reflexivity to analyze gender and online activism." *Journal of International Women's Studies*, 13(5), 31–45.

Nordås, R. and Davenport, C. (2013). "Fight the Youth: Youth Bulges and State Repression." *American Journal of Political Science*, 57(4), 926–940.

Olzak, S., Beasley, M., and Olivier, J. (2003). "The Impact of State Reforms on Protest Against Apartheid in South Africa." *Mobilization*, 8(1), 27–50.

Opp, K.-D. and Gern, C. (1993). "Dissident Groups, Personal Networks, and Spontaneous Cooperation: The East German Revolution of 1989." *American Sociological Review*, 58(5), 659–680.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). "Deep face recognition.." *BMVC*, Vol. 1, 6.

Pearlman, W. (2013). "Emotions and the Microfoundations of the Arab Uprisings." *Perspectives on Politics*, 11(02), 387–409.

Pfeffer, J. and Mayer, K. (2018). "Tampering with Twitter's Sample API." *EPJ Data Science*, 7(50), 1–21.

Purdy, C. (2018). "China is launching a dystopian program to monitor citizens in Beijing, <https://qz.com/1473966/china-is-starting-a-big-brother-monitoring-program-in-beijing/>.

Qin, B., Strömberg, D., and Wu, Y. (2017). "Why Does China Allow Freer Social Media? Protests Versus Surveillance and Propaganda." *Journal of Economic Perspectives*, 31(1), 117–140.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010). "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research*, 47(5), 651–660.

Rasler, K. (1996). "Concessions, Repression, and Political Protest in the Iranian Revolution." *American Sociological Review*, 61(1), 132–152.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*, 91–99.

Ritter, E. H. (2013). "Policy Disputes, Political Survival, and the Onset and Severity of State Repression." *Journal of Conflict Resolution*, 57(1), 1–26.

Ritter, E. H. and Conrad, C. R. (2016). "Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression." *American Political Science Review*, 110(1), 85–99.

Rizzo, H., Price, A. M., and Meyer, K. (2012). "Targeting Cultural Change in Repressive Environments: The Campaign against Sexual Harassment in Egypt." *Report No. 614*, Egyptian Center for Women's Rights, Cairo, <http://ecwronline.org/?p=1579>.

Robertson, G. B. (2007). "Strikes and Labor Organization in Hybrid Regimes." *American Political Science Review*, 101(04), 781–798.

Robnett, B. (1996). "African-American Women in the Civil Rights Movement, 1954-1965: Gender, Leadership, and Micromobilization." *American Journal of Sociology*, 101(6), 1661–1693.

Rosenblatt, F. (1958). "The perceptron: a probabilistic model for information storage and organization in the brain.." *Psychological review*, 65(6), 386.

Salehyan, I., Hendrix, C., Hammer, J., Case, C., Linebarger, C., Stull, E., and Williams, J. (2012). "Social Conflict in Africa: A New Database." *International Interactions*, 38(4), 503–511.

Schaftenaar, S. (2017). "How (wo)men rebel: Exploring the effect of gender equality on nonviolent and armed conflict onset." *Journal of Peace Research*, 54(6), 762–776.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Schweingruber, D. and McPhail, C. (1999). "A Method for Systematically Observing and Recording Collective Action." *Sociological Methods & Research*, 27(4), 451–498.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2016). "Grad-cam: Visual explanations from deep networks via gradient-based localization." *See https://arxiv. org/abs/1610.02391 v3*, 7(8).

Shaban, H. (2018). "Amazon employees demand company cut ties with ICE, <https://www.washingtonpost.com/news/the-switch/wp/2018/06/22/amazon-employees-demand-company-cut-ties-with-ice> (jun).

Shakir, O. (2014). "All According to Plan: The Rab'a Massacre and Mass Killings of Protesters in Egypt." *Report no.*, Human Rights Watch.

Shellman, S. M., Levey, B. P., and Young, J. K. (2013). "Shifting sands: Explaining and predicting phase shifts by dissident organizations." *Journal of Peace Research*, 50(3), 319–336.

Siegel, D. A. (2011). "When Does Repression Work? Collective Action in Social Networks." *The Journal of Politics*, 73(04), 993–1010.

Sloan, L. and Morgan, J. (2015). "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter." *PLoS ONE*, 10(11), 1–15.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). "Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter." *Sociological Research Online*, 18(3), 1–15.

Sobolev, A., Joo, J., Chen, K., and Steinert-Threlkeld, Z. C. (2019). "Newspapers and Social Media Accurately Measure Protest Size." *Working paper*.

Steinert-Threlkeld, Z. C. (2017). "Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring." *American Political Science Review*, 111(02), 379–403.

Steinert-Threlkeld, Z. C. (2018). *Twitter as Data*. Cambridge University Press.

Stephan, M. J. and Chenoweth, E. (2008). "Why Civil Resistance Works." *International Security*, 33(1), 7–44.

Sullivan, C. M. (2016). "Political Repression and the Destruction of Dissident Organizations." *World Politics*, 68(4), 645–676.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). "Deep learning face representation by joint identification-verification." *Advances in neural information processing systems*, 1988–1996.

Sutton, J., Butcher, C. R., and Svensson, I. (2014). "Explaining political jiu-jitsu: Institution-building and the outcomes of regime violence against unarmed protests." *Journal of Peace Research*, 51(5), 559–573.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). "YFCC100M: The New Data in Multimedia Research." *Communications of the ACM*, 64–73.

Tilly, C. and Wood, L. J. (2012). *Social Movements, 1768-2012*. Paradigm Publishers, 3rd edition.

Torres, M. (2018). "Give me the full picture: Using computer vision to understand visual frames and political communication." Working paper.

Tucker, J. A. (2007). "Enough! Electoral Fraud, Collective Action Problems, and Post-Communist Colored Revolutions." *Perspectives on Politics*, 5(03), 535.

Tufekci, Z. (2014). "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls Pre-print." *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, Ann Arbor.

Urdal, H. (2006). "A Clash of Generations? Youth Bulges and Political Violence." *International Studies Quarterly*, 50, 607–629.

Urdal, H. and Hoelscher, K. (2012). "Explaining Urban Social Disorder and Violence: An Empirical Study of Event Data from Asian and Sub-Saharan African Cities." *International Interactions*, 38(4), 512–528.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). "Online Human-Bot Interactions: Detection, Estimation, and Characterization." *Working paper*.

Weidmann, N. B. (2014). "On the Accuracy of Media-based Conflict Event Data." *Journal of Conflict Resolution*, 59(6), 1129–1149.

Weyland, K. (2012). "The Arab Spring: Why the Surprising Similarities with the Revolutionary Wave of 1848?." *Perspectives on Politics*, 10(04), 917–934.

White, M. (2016). *The End of Protest: A New Playbook for Revolution*. Knopf Canada.

Wilson, R. E., Gosling, S. D., and Graham, L. T. (2012). "A Review of Facebook Research in the Social Sciences." *Perspectives on Psychological Science*, 7(3), 203–220.

Won, D., Steinert-Threlkeld, Z. C., and Joo, J. (2017). "Protest activity detection and perceived violence estimation from social media images." *Proceedings of the 2017 ACM on Multimedia Conference*, ACM, 786–794.

Xu, H., Gao, Y., Yu, F., and Darrell, T. (2017). "End-to-end learning of driving models from large-scale video datasets." *arXiv preprint*.

Young, L. E. (2019). "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review*, 113(1), 140–155.

Zhang, H. and Pan, J. (2019). "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology*, 49, 1–48.

Zhao, D. (1998). "Ecologies of Social Movements: Student Mobilization during the 1989 Prodemocracy Movement in Beijing." *American Journal of Sociology*, 103(6), 1493–1529.

Supplementary Materials for

How Violence Affects Protests

APPENDIX S1. DETAIL

S1.1 Convolutional Neural Networks

We use convolutional neural networks (CNN) to identify and analyze protest images. A CNN is a type of artificial neural network, a machine learning algorithm inspired by the human brain (Rosenblatt, 1958), that has gained widespread adoption in the field of computer vision. It has been successful in various applications including face recognition (Sun et al., 2014; Parkhi et al., 2015; Baltrušaitis et al., 2016), object detection (Girshick et al., 2014; Ren et al., 2015; Redmon et al., 2016), and self-driving cars (Huval et al., 2015; Bojarski et al., 2016; Xu et al., 2017). For methodological detail on computer vision for political scientists, see Joo and Steinert-Threlkeld (2018).

A CNN is a function whose outputs are computed through a series of sequential operations from the input values. For example, in image classification, the input is an image (i.e., an array of color intensities at pixels) and the output is the class that the image belongs to, e.g., an object category such as "police" or "male". A CNN transforms the given input through many operations until it reaches the final step which produces the output. Each operation is also called a layer, and a CNN usually has multiple "convolutional" layers. A convolutional layer performs convolution, which consists of an element-wise multiplication between pixel[13] values and filter values (connection strengths) and a summation over adjacent pixels: this essentially measures how well the appearance of an input image matches the "template" that the model learned in training. A CNN, as well as other artificial neural networks, is trained to minimize a loss function, a measure of difference between the model prediction and ground truth label. This optimization is typically done by stochastic gradient descent.

[13] A CNN contains many layers and the output of a layer becomes the input of the next layer. The input to the first layer is the original input image's pixel intensities. For non-first layers, their inputs are given from nodes on two dimensional grid in the previous layer, not from the image pixels.

Each CNN is defined by its architecture – the structural configuration specifying the number of layers, the order of their placement, and the types of non-linear transformations used. There exist many different CNN architectures with different properties. The architecture of our model is a "Residual Network" (ResNet) (He et al., 2016b) and has 50 convolutional layers. ResNet has been used in many of the state-of-the-art computer vision applications such as object detection (Ren et al., 2015) and human pose estimation (Güler et al., 2018). We use a ResNet model pre-trained on ImageNet data and finetune it with our data.

This paper does not use tweet text because they do not measure violence, identity, or free riding as precisely as images. Decades of construction of event data, via hand and computer coding, has not been able to generate a measure of state or protester violence more refined than an ordinal measure; images allow for violence to be measured as a continuous variable. Measuring cleavages from text requires knowing the identity of accounts and would require orders of magnitude more user data; this exercise would not produce time varying measures because they would be about the account, not other protesters (Mislove et al., 2011; Sloan et al., 2013). Event datasets that measure cleavage spanning use newspaper text, which often does not report protester demographic information, and so measures are fixed at the movement level (Heaney and Rojas, 2008; Kern, 2011; Wilson et al., 2012; Fisher et al., 2017); images permit the measurement of the mass of protesters and their daily change. Measuring free riding from tweet text would require building a classifier, for each language in our dataset, for specific statements such as "I am not going to protest because it will not make a difference"; images that can induce free riding are easier to identify than specific tweets because visual language is universal (Graber, 1996).

S1.2 Classifier Calibration

For binary variables in our analysis, we need to transform continuous outputs from CNN to binary values (0 or 1) by choosing a decision threshold such that we can determine if an image contains the variable of interest. The optimal decision threshold needs to be chosen so that it can balance good true positive and true negative rates, evaluated on the target data distribution (i.e., not the distribution in our development set). To this end, we chose 3,000 protest images from additional random samples from our Twitter pipeline and used Amazon Mechanical Turk to annotate them. We then generated a precision-recall curve for each attribute and chose the threshold at the minimum precision of .85.[14] For each image and each attribute, our model therefore produces a probability estimate (a real number) via the classifier as well as a binary output (0 or 1). Figure A2 shows the precision-recall curve for each attribute, providing the threshold value for each. The twelve attributes and their thresholds are shown in Table A9.

[14]One could also use another method such as F-measure to choose the optimal decision threshold. In our study, it is more important to maintain the minimum precision (true positive rate) at a high point for every attribute, rather than trying to detect more relevant images while making more mistakes.

Table A8. List of visual attributes.

Attribute	Description	Hypothesis	Expectation
1. Protester Violence	How violent protesters are.	Violence (H1)	–
2. State Violence	How violent the state is.	Violence (H1)	+, –
3. Police	Police or troops are present in the scene.	Violence (H1)	–
4. Fire	There is fire or smoke in the scene.	Violence (H1)	–
5. **Gender**	**Is the face male or female?**	**Cleavages (H2)**	+, –
6. **Race**	**Is the face White, Middle Eastern, East Asian, Southeast Asian, Black, Indian, or Latino?**	**Cleavages (H2)**	+, –
7. **Age**	**0-2, 3-9, 10-19, [...], 70**	**Cleavages (H2)**	+, –
8. **Face**	**Presence of a face.**	Protest size	
9. Group 20	There are roughly more than 20 people in the scene.	Future: free riding	
10. Group 100	There are roughly more than 100 people in the scene.	Future: free riding	
## Children	Children are in the scene.	N/A	N/A
## Shout	One or more people shouting.	N/A	N/A
## Photo	Protesters holding signs or a photograph of a person (politicians or celebrities).	N/A	N/A
## Flag	There are flags in the scene.	N/A	N/A
## Night	It is at night.	N/A	N/A
## Sign	Protesters holding a visual sign (on paper, panel, or wood).	N/A	N/A

NB: Attributes without numbers could not be classified precisely enough to be included in regression models.
NB: Attributes in **bold** are generated using the face classifier.

Table A9. Attributes and Thresholds

Attribute	Threshold
Protester Violence	.021
State Violence	.01
Police	.937
Fire	.37
Child	.15
Small Group	.725
Large Group	.509
Shout	.355
Photo	.815
Flag	.187
Night	.359
Sign	.744

Fig. A1. Examples of Our Annotation Interface (in Amazon Mechanical Turk)

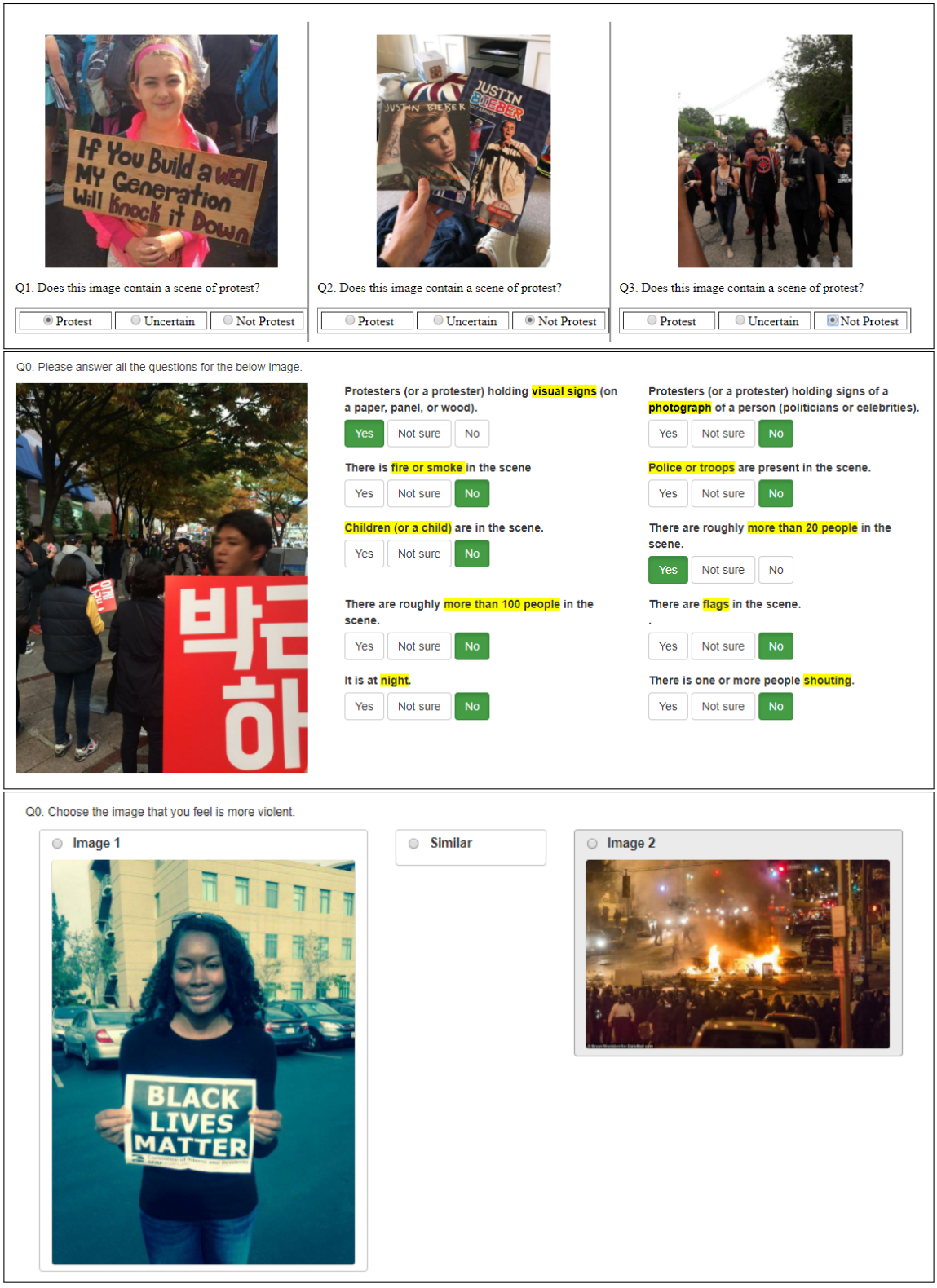


Fig. A2. Precision-Recall Curves For Binary Attributes. AP stands for average precision, which is the standard accuracy measure for binary classification. AP is also equal to the area under the precision-recall curve.

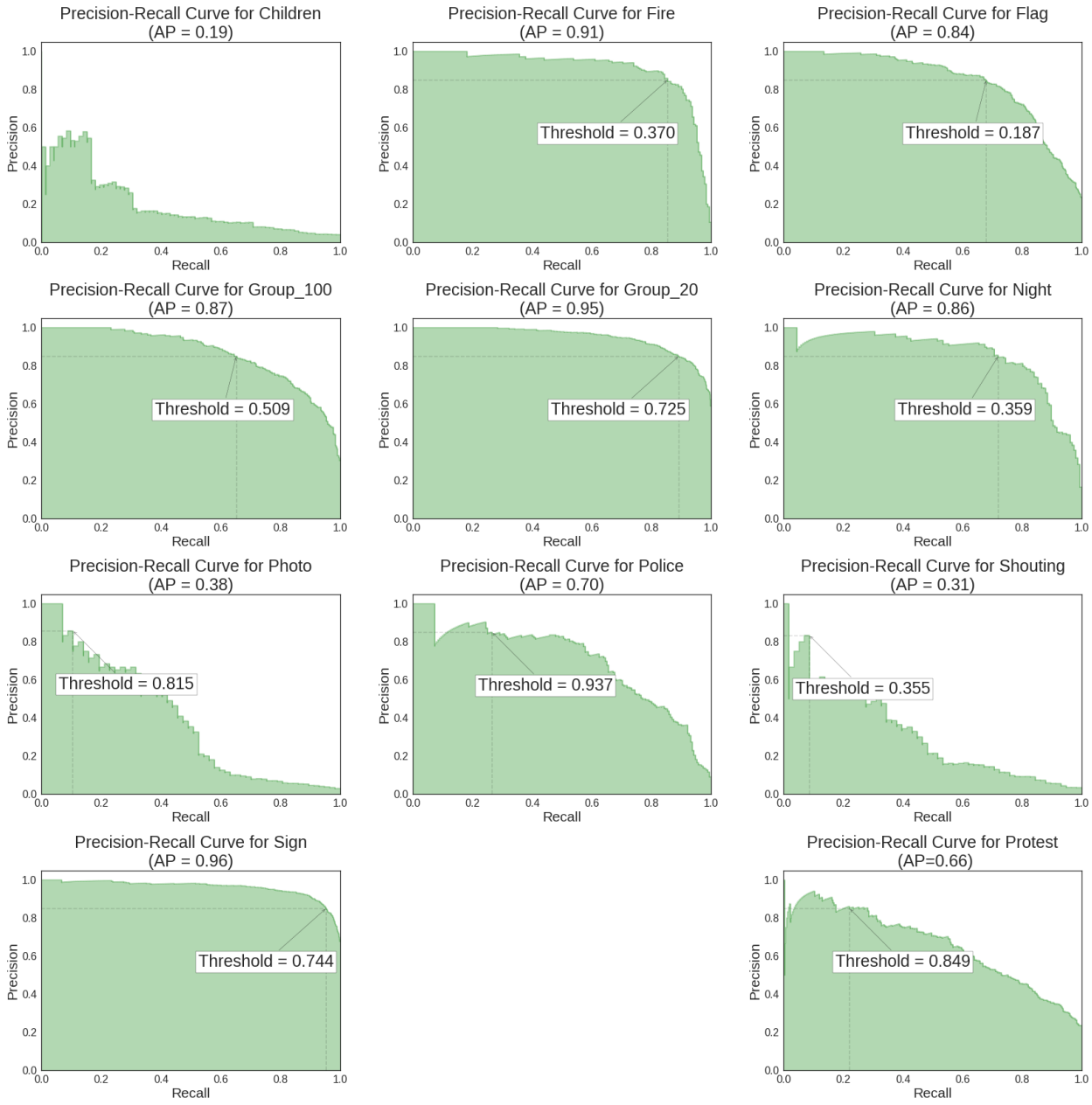


Fig. A3. Example Results of Our Face Model

S1.3 Evaluating the CNN

Figure A4 shows the model performance measured on the validation set. The Receiver-Operator Curve (ROC) documents the relationship between false-positive and true-positives, with a higher area-under-curve (AUC) corresponding to a better accuracy. Visually, the closer the curve is to the upper-left corner, the better the classifier for that label.

Figure A5 shows a scatterplot of the classifier's output for violence against the rating recovered from the Bradley-Terry model. It also shows the ROC curve for protester violence and state violence. All three subfigures demonstrate strong performance of our classifier's ability to measure perceived violence.

To intuitively visualize how the classifier works, we use Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2016). Grad-CAM highlights important regions for classifying the concept in an image. Grad-CAM highlights such regions by tracing back the classification outcome to the input image through passing gradients. The results are shown in Figure A6, with red color indicating more important regions. For technical details, see Selvaraju et al. (2016).

Figure A7 arrays images from each category by the classification scores from the CNN. As the classification score approaches 1 for each category, images more closely exhibit the visual concept.

Fig. A4. Model Performance

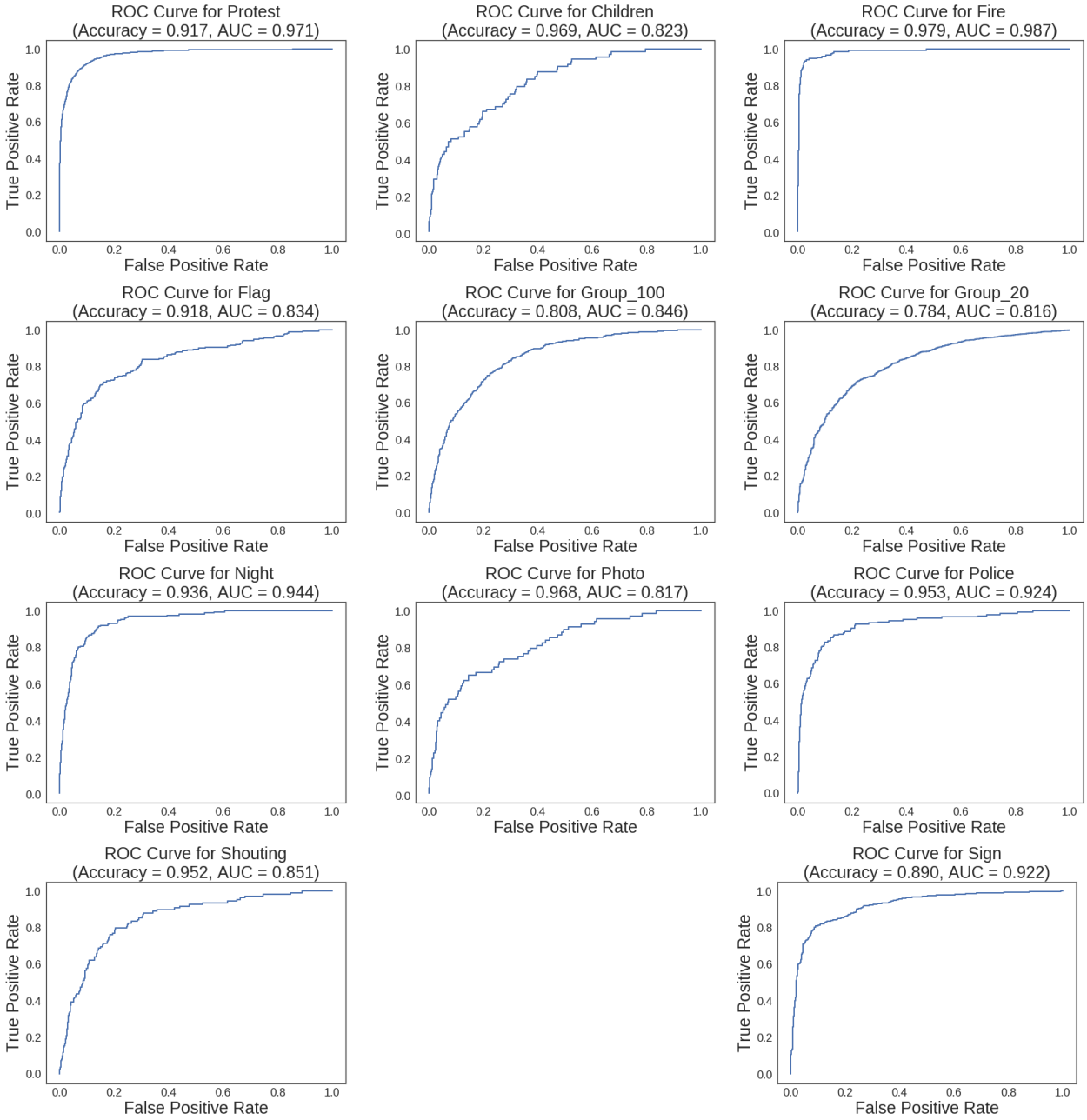


Fig. A5. Validating Violence Measurement

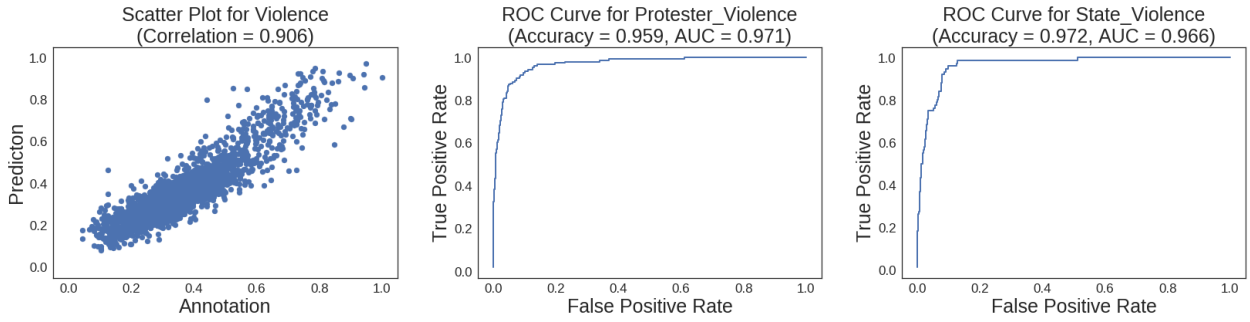


Fig. A6. Visualization of Region Importance in Classification Using Grad-CAM: Important regions which more contribute to the classification for each attribute are highlighted in red.

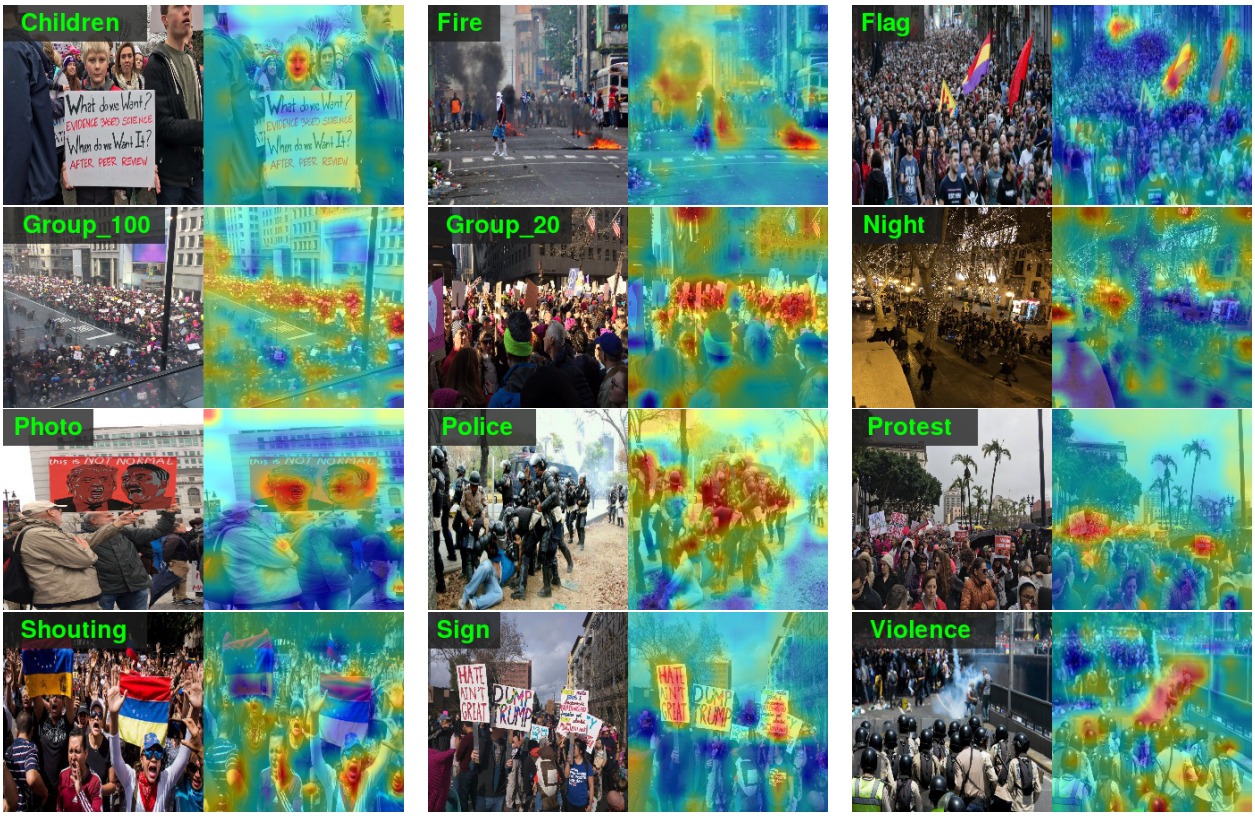


Fig. A7. Sample Classifier Estimates by Category: Images are ordered by their classification scores. (Blue lines mark the exact classification scores of corresponding images)

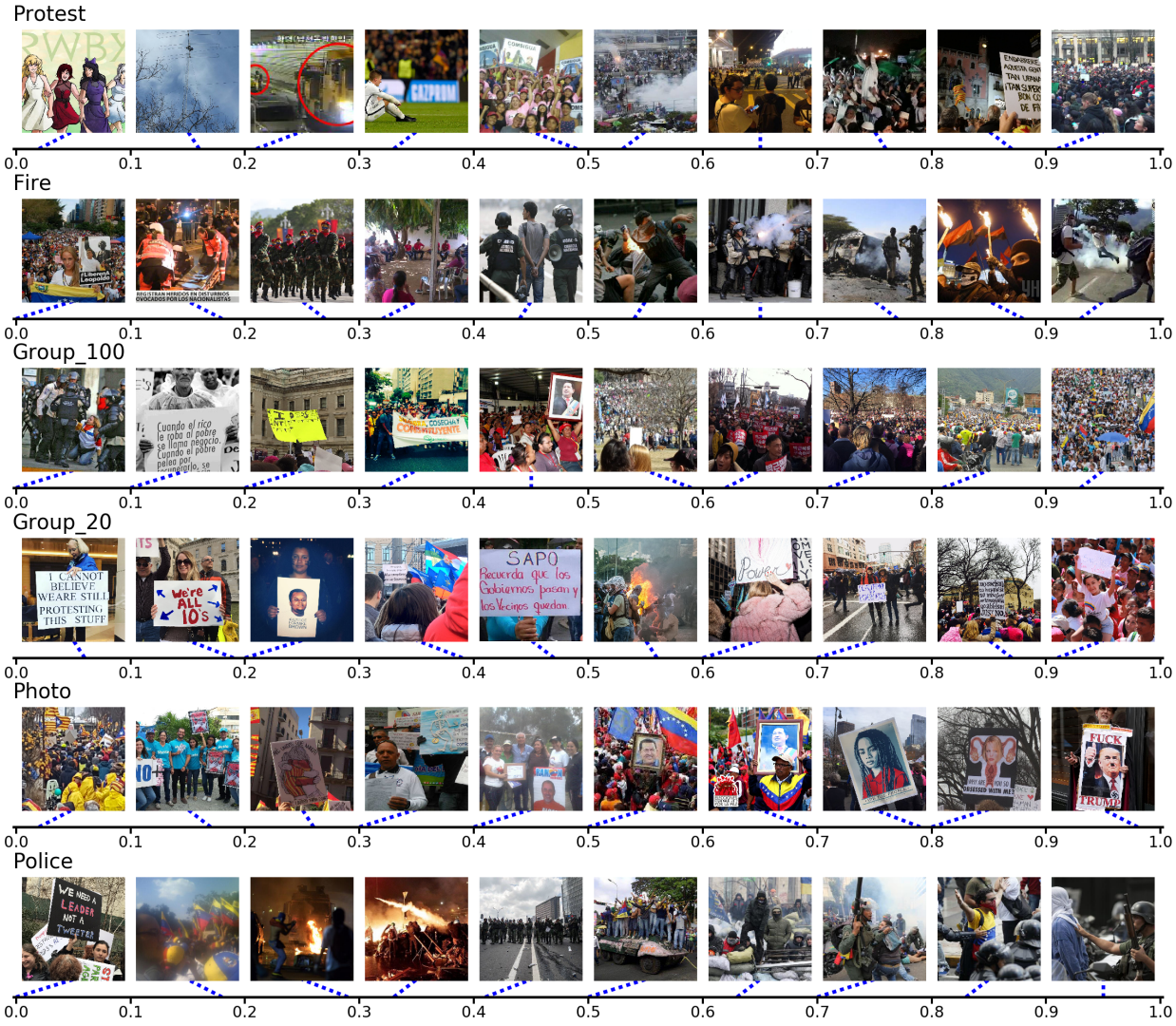
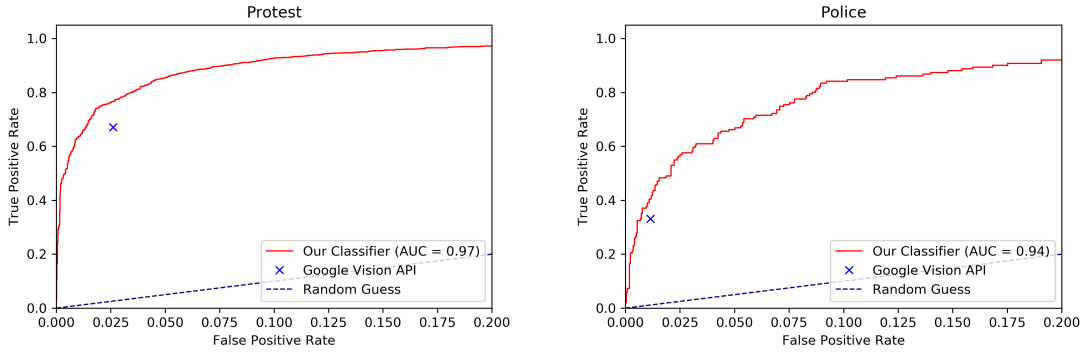


Fig. A7. Sample Classifier Estimates by Category (Continued)

Finally, to compare the classification performance of existing commercial classifiers against our own classifier, we used Google Vision's label detection on the test set of UCLA Protest Image Database (Won et al., 2017) and measured the classification accuracy. This dataset has 11,000 test images with various labels related to protest activity such as the presence of protesters or police officers in images. Since Google's label detection automatically identifies visual concepts and objects in many categories, including protest and police, from an input image, we directly compared its accuracy with our model accuracy. As shown in Figure A8, the protest and scene models classified protest and police more accurately than the Google Vision API. The superior result is most likely due to the fact that we specifically collected diverse protest images and hard-negatives (i.e., non-protest images which look like protest) from many sources. The Google Vision API may perform better on general image classification and can be very useful when one does not have any training data.

Fig. A8. Classification performance comparison between our model and the public model from Google's Vision API.



APPENDIX S2. BIAS

In the United States, Twitter users who geotag are richer, more likely to live in cities, young, and non-white (Malik et al., 2015). In the United Kingdom, Twitter users are younger, more educated, more likely to be male, and more politically engaged (but less likely to vote) than others (Mellon and Prosser, 2017). Once on Twitter, geotagging users are slightly older than non-geotaggers, there is some difference in rates of geotagging across profession, and there is large variation by tweet language in the percentage of users who geotag (a low of 0.4% for Arabic accounts to a high of 8.3% for Turkish, with an average of 3.1%) (Sloan and Morgan, 2015).

Though Twitter users differ from non-Twitter users and those who assign locations to their tweets differ from those who do not, there is no *a priori* reason to expect that they differ in the type of protest images they share. Conditional on being at a protest, there is no reason to think that the contents of a geotagged protest image should systematically differ from a protest image that is not geotagged. Of anything, the importance of social media for tactical coordination of protests means that geotagged tweets should be *more* likely to represent a protest than one that is not (Gunitsky, 2015; Little, 2015).

Section S6 compares users who tweet protest images to those who tweet non-protest images. More accounts share protest images than non-protest images, and they have fewer followers than those tweeting non-protest images. There is no statistically significant difference in the account age or frequency of tweeting. While this comparison does not prove that the protest images are an unbiased representation of the protest itself, it at least appears to be the case that the accounts themselves do not appear to be any more biased, and are probably actually more representative, than the larger Twittersphere. The protest photos appear to come from more “normal" users than those who normally tweet images.

If bias in protest data from geolocated images shared on Twitter exists, it should nonetheless be less than that which exists from relying on any text that is not a police archive. The main source of information for existing event data, newspapers, have large, well-known biases that result from incentives that are much weaker on social media. Newspapers are much more likely to cover large

events than small ones (Mccarthy et al., 1996) as well as events perceived to be of interest to their subscribers (Myers and Caniglia, 2004; Baum and Zhukov, 2015). Events away from urban centers are less likely to receive coverage (Kalyvas, 2004; Weidmann, 2014), as are ones that are parts of a larger wave of events (Hellmeier et al., 2018). Given the increasing consolidation of the newspaper industry, these biases are likely to have become more consistent across sources (Baum and Zhukov, 2018).

These biases exist because newspapers have to maximize readership (advertising, newsstand, and subscription revenue) while constrained by space. This constraint puts an emphasis on reporting novel or unexpected events such as violent attacks or large protests. Even if readership is national - and most newspapers have local or, at best, regional circulation - events are still selected based on their appeal to the readers. The need to daily attract readers means coverage of events quickly tapers regardless of the event duration (Hellmeier et al., 2018). For a more extensive explanation of bias in news coverage, see Earl et al. (2004).

Social media platforms do not face these same pressures. While their business model is more focused on attracting eyeballs than newspapers are, because they do not have subscribers or newsstand sales, there is essentially zero restriction on the space in which to publish.[15] Whether or not the platforms, such as Twitter or Facebook, should be treated as media companies is a separate issue, but one way in which they are not like other media is that they do not employ people to create the information featured on their platform, the way newspapers pay journalists. Given the essentially infinite supply of posts and the lack of control over content providers, there should therefore be much less selection pressure on what appears on social media.[16] Because newspapers face scarcity constraints that social media do not, the latter should be much more likely to provide a less biased account of the world than newspapers. In providing orders of magnitude more posts than newspapers do articles, social media are closer in scope to government archives than they are

[15]Each new post imposes a marginal cost - server space and electricity - on the platform that is much smaller than article for a newspaper.

[16]Social media platforms increase user engagement by selectively presenting posts to users. While this algorithmic process may present users with biased interpretations of events, that process is not used to decide which tweets to send to the API (Pfeffer and Mayer, 2018).

newspapers (Sullivan, 2016). See Sobolev et al. (2019) for a more extensive comparison of bias in newspaper and social media event data.

APPENDIX S3. COUNTRY PERIODS ANALYZED

Table A10. Protest Periods

Country	Start	End	Issue	Tweets
Images	Protest Images/Day			
Belarus	02.18.2017	05.02.2017	Unemployment Tax	2.18
Burundi	04.01.2015	12.01.2015	Elections	.06
Cameroon	11.01.2016	12.01.2017	Bilingualism	.06
Egypt	06.01.2017	06.31.2017	Islands to Saudi Arabia	5.38
Gabon	08.20.2016	09.27.2016	Elections	.235
Hong Kong	**2014.09.18**	**2014.12.23**	**China Reforms**	**5.82**
Pakistan	11.01.2017	11.30.2017	Blasphemy protests	.941
Russia	***03.12.2017***	***04.26.2017***	***Corruption***	***19.3***
Catalonia, Spain	***2017.09.01***	***2017.12.31***	***Secession***	***31.8***
South Korea	**2016.10.20**	**2017.03.14**	**Anti-incumbency**	**8.04**
Togo	08.01.2017	12.01.2017	Anti-incumbency	.23
Ukraine	11.21.2013	03.21.2014	European Integration	3.32
United States	2017.01.20	2017.01.22	Women's March	9,034.33
Venezuela	**2014.03.27**	**2015.02.08**	**Grievances**	**31.20**
Venezuela	**12.29.2016**	**12.17.2017**	**Anti-Maduro**	**16.40**

APPENDIX S4. CITY-DAY CORRELATION

Figure A9 shows the correlation between the models' variables. The only variables with correlation above .8 are the two group variables. These correlations are higher than the per tweet ones: the correlation comes from aggregating different photos to the city-day level, not from the classifier producing similar estimates for different labels. The per tweet correlation is shown in Figure A10.

APPENDIX S5. TWEET LEVEL CORRELATION

Figure A10 shows the correlation between variables at the image level. The only correlation above .8 is male faces and white faces with the number of faces. These correlations are lower than the country-day correlations, meaning that correlations in the aggregated data come from multiple mechanisms occurring during a protest, not noise in the classification of individual photographs.

Fig. A9. Covariate Correlation, by City-day

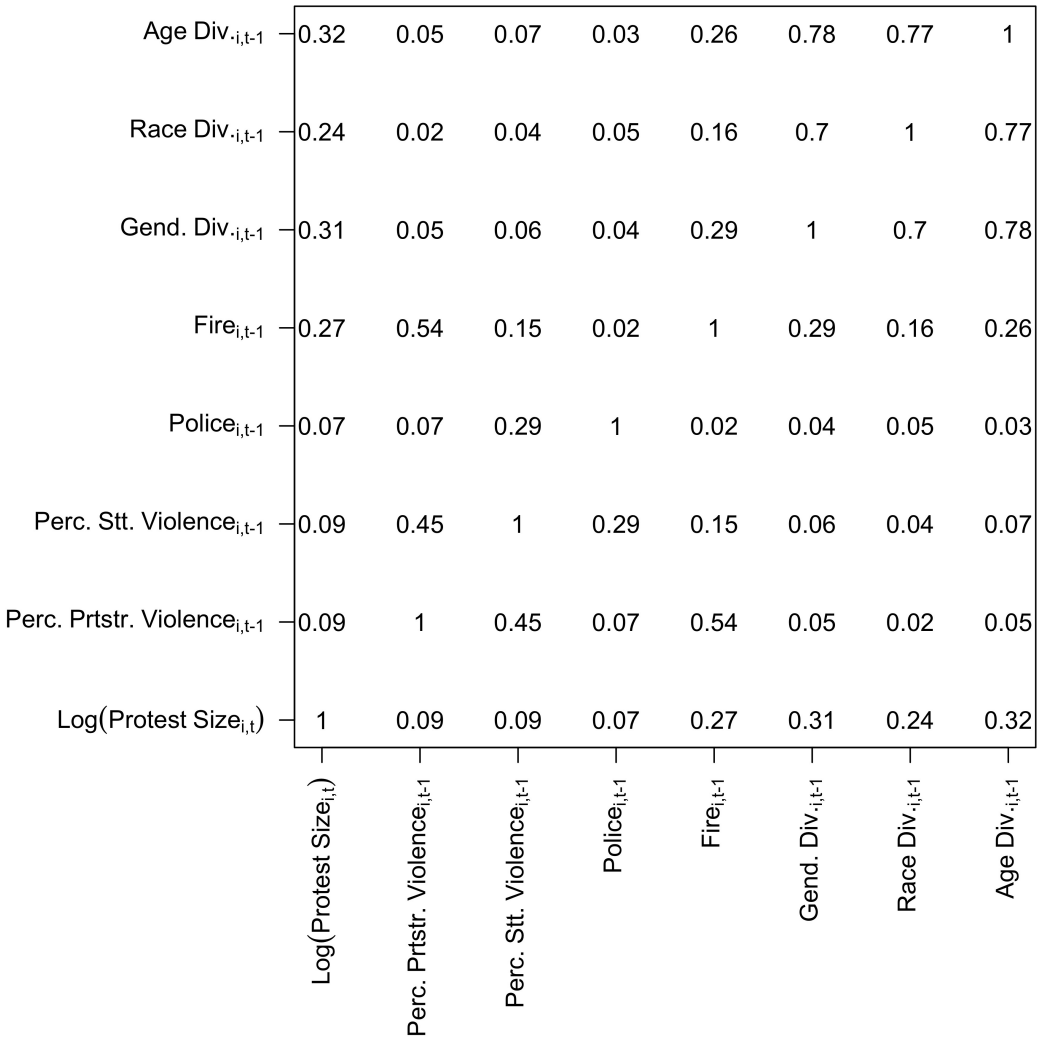
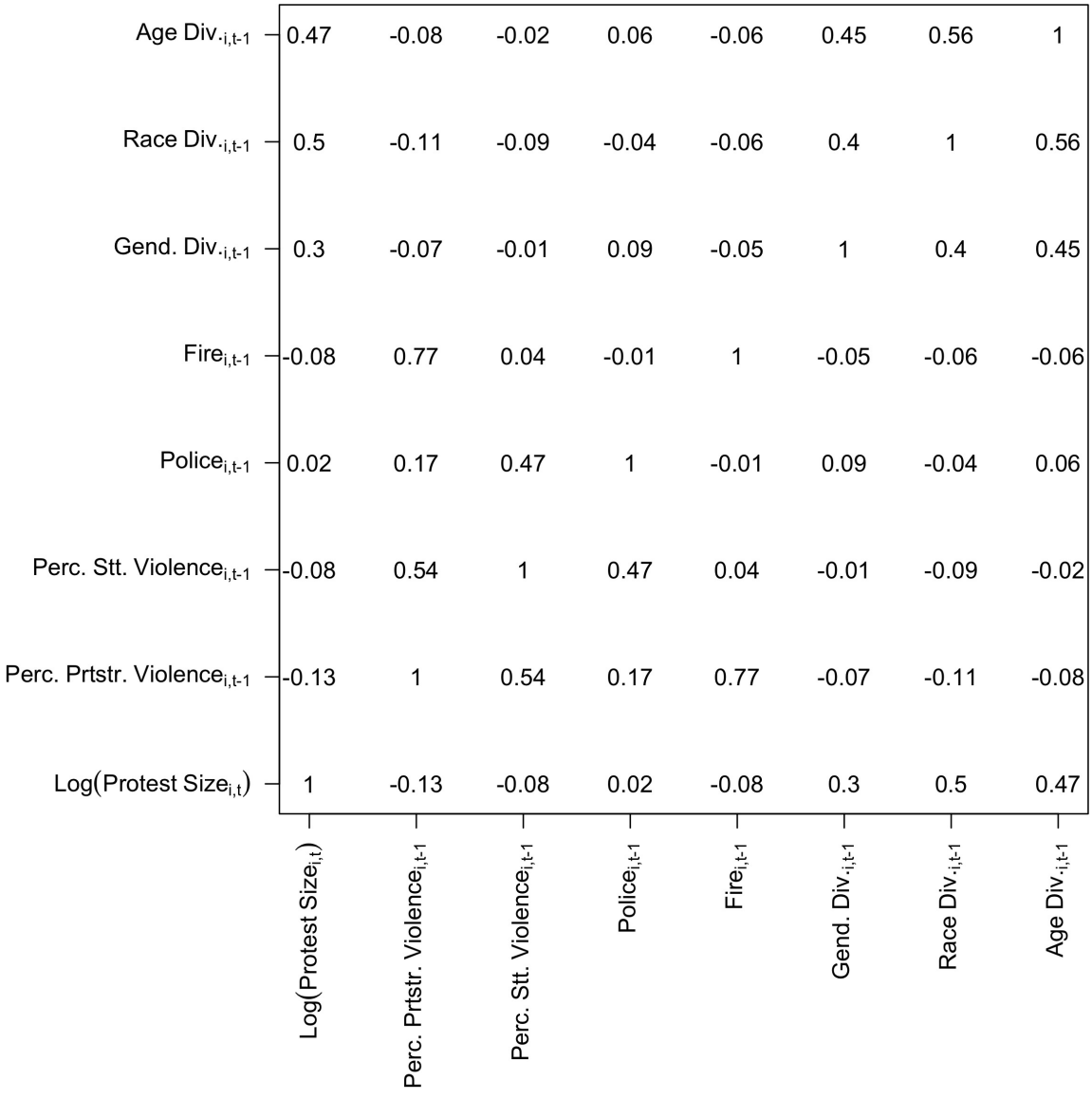


Fig. A10. Covariate Correlation, by Tweet



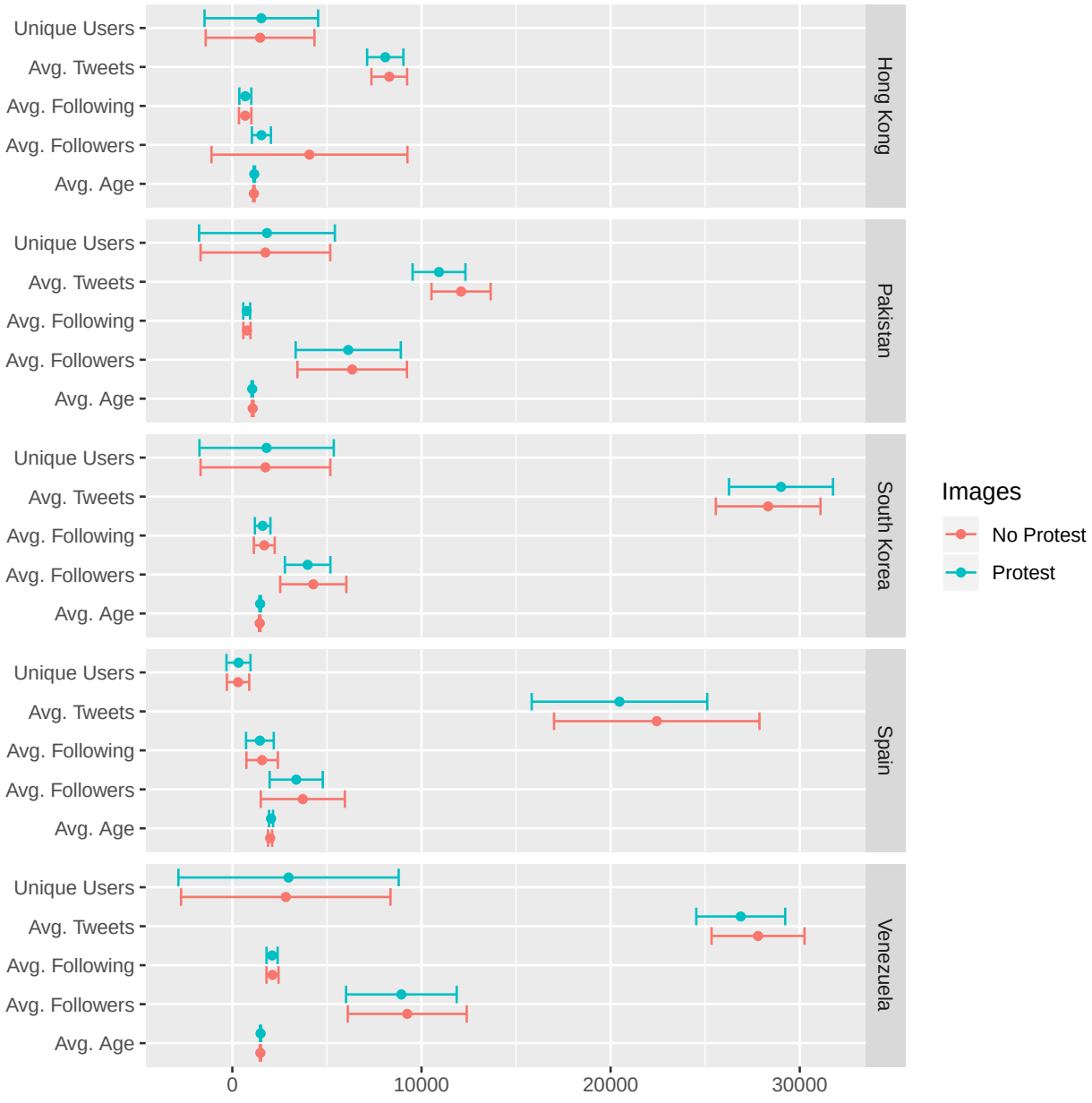
APPENDIX S6. COMPARING IMAGE SHARERS TO PROTEST IMAGE SHARERS

Interested in whether or not those who tweet images of protests differ from those who tweet images, we analyzed account covariates by country-photo type (protest or not protest). For each set of photos, we kept each user only once, which discarded about 40% of tweets in each set. For users who tweeted the same image type multiple times, we randomly keep one tweet, and an account can appear in both samples. The account characteristics we analyze are the number of followers, following, statuses, and account age (days on Twitter). We also count the number of unique users in each country by image type. The points estimates and 95% confidence intervals are shown in Figure A11.

More, less popular people share protest images than non-protest images. In all countries, there are more accounts that tweet protest images than non-protest images, though the difference is only statistically significant in Venezuela. These extra accounts have fewer followers on average than those tweeting non-protest images, though the difference is not statistically significant in Ukraine.

The two sets of users do not differ in how engaged they are on Twitter. In no country do the protest and non-protest accounts follow different numbers of users. They have each been on Twitter the same amount of time. (In Russia only, these accounts have been on Twitter longer, an average of 43.21 days more.) In Venezuela, the protest users are slightly less active on Twitter, but in Russia they are more active; there is no statistically significant difference for Ukraine.

Fig. A11. Users Tweeting Protests vs. Not



APPENDIX S7. VALIDATING THE SCENE AND FACE CLASSIFIERS

Figure A12 shows the distribution of protester violence by country. Figure A13 shows the distribution of race entropy by country. Figure A14 shows the distribution of age entropy by country.

Fig. A12. Distribution of Protester Violence by Country

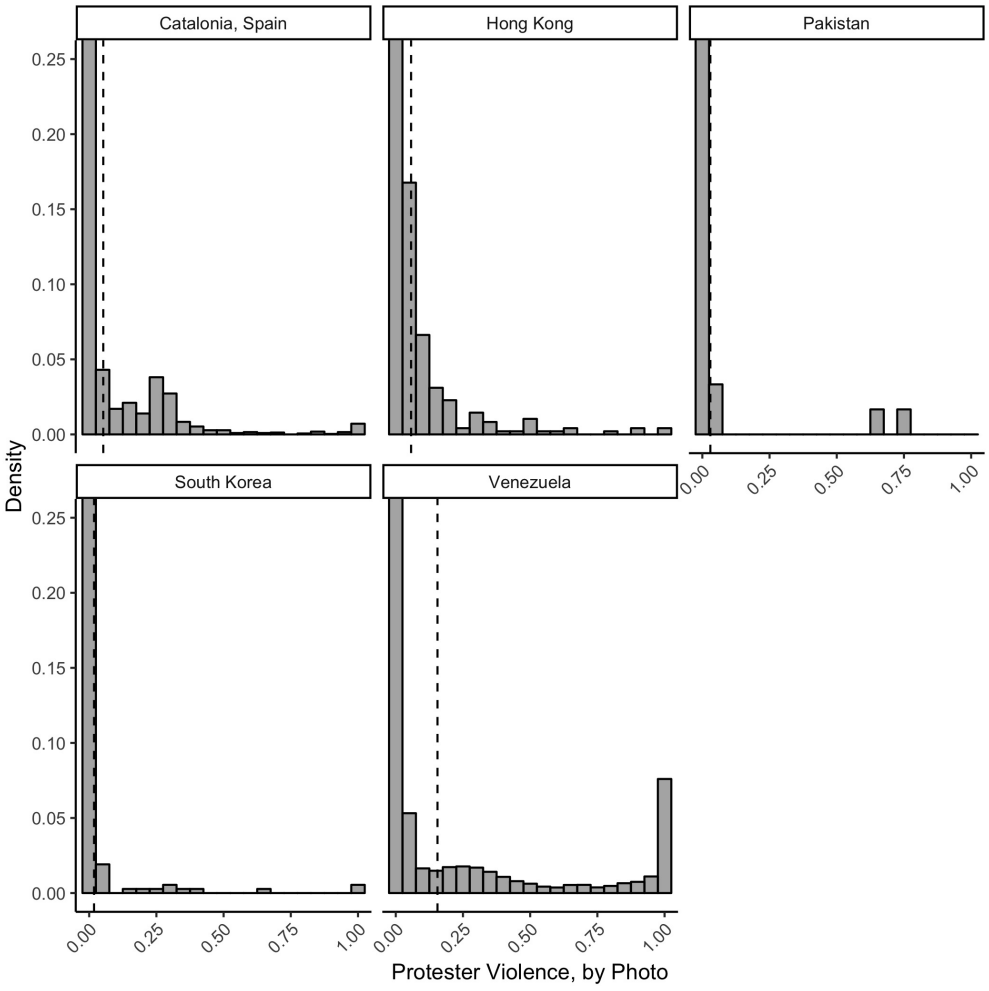


Fig. A13. Distribution of Race Entropy by Country

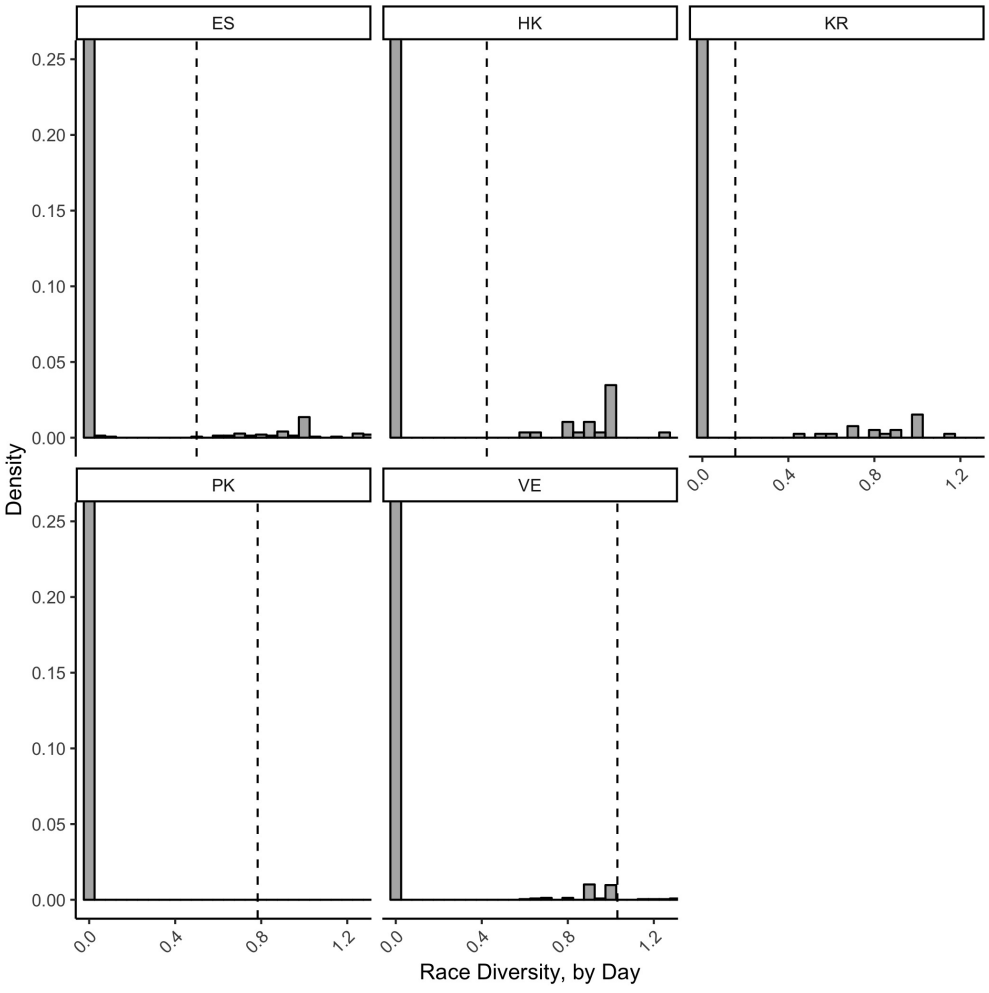
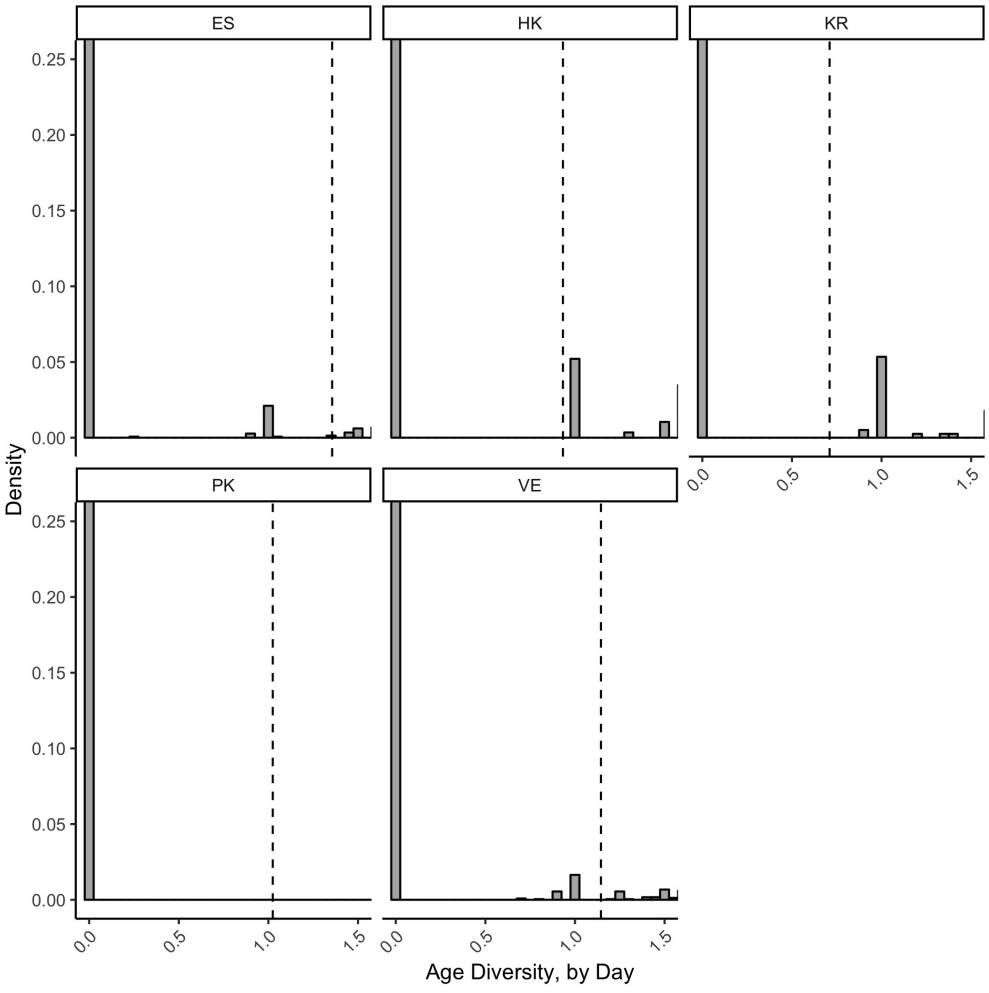


Fig. A14. Distribution of Age Entropy by Country



APPENDIX S8. ADDITIONAL ROBUSTNESS CHECKS

S8.1 Bot Distribution

Table A11 compliments Table 7. No more than 6.7% of accounts are from bots, and no more than 6.5% of tweets.

Table A11. Distribution of Bots by City

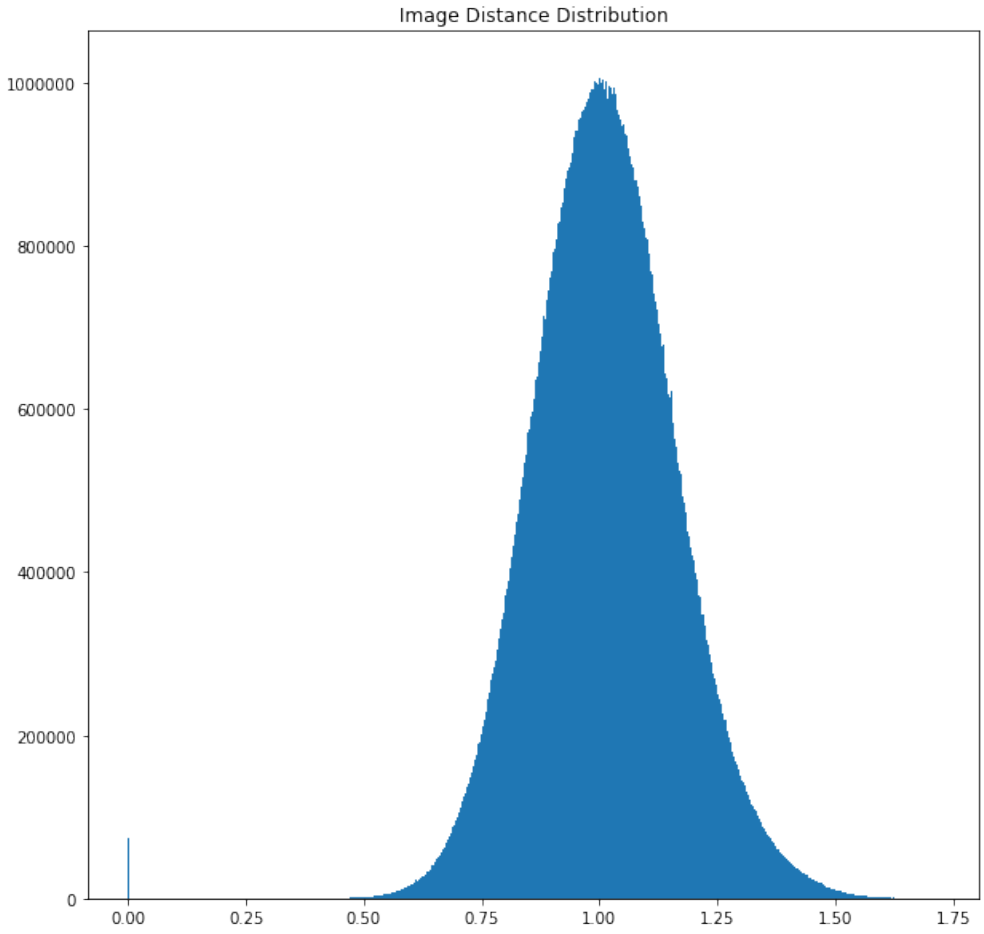
City	Avg. Bot Score	Max. Bot Score	SD Bot Score	Percent Tweets from Bots	Percent Accounts of Bots
Ciutat Vella	0.131	0.948	0.297	0.108	0.040
Lahore	0.067	0.559	0.173	0.100	0.143
Sant Salvador de Guardiola	0.066	0.611	0.155	0.083	0.094
Granera	0.053	0.812	0.156	0.054	0.065
Valencia	0.063	0.921	0.145	0.052	0.058
Tarragona	0.058	0.829	0.172	0.052	0.079
Central	0.050	0.845	0.156	0.051	0.133
Maracaibo	0.050	0.685	0.115	0.039	0.037
Seoul	0.145	0.905	0.170	0.035	0.056
Barcelona	0.029	0.939	0.110	0.024	0.022
Caucagua	0.054	0.905	0.118	0.020	0.027
Boca del Rio	0.032	0.829	0.117	0.020	0.047
Girona	0.034	0.637	0.088	0.019	0.038
Sant Feliu de Pallerols	0.040	0.661	0.086	0.018	0.048
Caracas	0.043	0.942	0.103	0.010	0.025
Granollers	0.011	0.084	0.019	0	0
Kimhae	0.052	0.054	0.009	0	0
Kowloon	0.021	0.355	0.062	0	0
Lleida	0.018	0.355	0.058	0	0
Mataró	0.007	0.054	0.009	0	0
Reus	0.018	0.297	0.051	0	0
Sabadell	0.018	0.355	0.044	0	0
Sant Cugat del Vallè	0.015	0.270	0.047	0	0
Terrassa	0.005	0.030	0.006	0	0

S8.2 Deduplicating Images

To deduplicate images, we extracted 1,000 features from a pre-trained ResNet50 model (He et al., 2016a). Conventional image preprocessing methods for deep learning models were used. Each image was resized to 256 x 256 pixels. Then, a center-crop of 224 x 224 pixels was performed. Finally, the cropped images were normalized to the mean and standard deviation of the ImageNet dataset (Deng et al., 2009). The 1,000 feature vector of each sample was normalized to unit norm. The L2 distance among the normalized data is computed, and images are considered matches if the

distance is less than a threshold of 0.2. The histogram of the distribution of distances in shown in Figure A15.

Fig. A15. Distribution of Pairwise Image Distances



Two manual checks verify these results. The largest 90 clusters were manually inspected and no images were misidentified as duplicates. The 220 most common images identified as duplicates, shared 2,500 times, were inspected, and none were misidentified as duplicates.

Table A12 shows the percentage of tweets per city that are duplicates.

Figure A16 shows the number of times each image appears in the dataset.

Figure A17 shows the distribution of the percentage of images per city that are duplicates. Except for Kimhae, the spike at 1 is cities with 1 image.

Table A12. Duplicate Images

City	Percentage Duplicates	Total Tweets
Kimhae	0.957	46
Sant Feliu de Pallerols	0.621	58
Girona	0.500	108
Caracas	0.463	2, 105
Mataró	0.425	40
Sant Cugat del Vallè	0.424	33
Caucagua	0.281	224
Tarragona	0.274	62
Valencia	0.269	167
Sant Salvador de Guardiola	0.250	40
Maracaibo	0.243	152
Terrassa	0.237	59
Boca del Rio	0.227	66
Sabadell	0.187	75
Reus	0.184	38
Barcelona	0.182	1, 338
Granera	0.154	65
Granollers	0.150	20
Lleida	0.116	43
Ciutat Vella	0.100	40
Seoul	0.052	326
Central	0	41
Kowloon	0	66
Lahore	0	5

Fig. A16. Distribution of Number of Duplicates

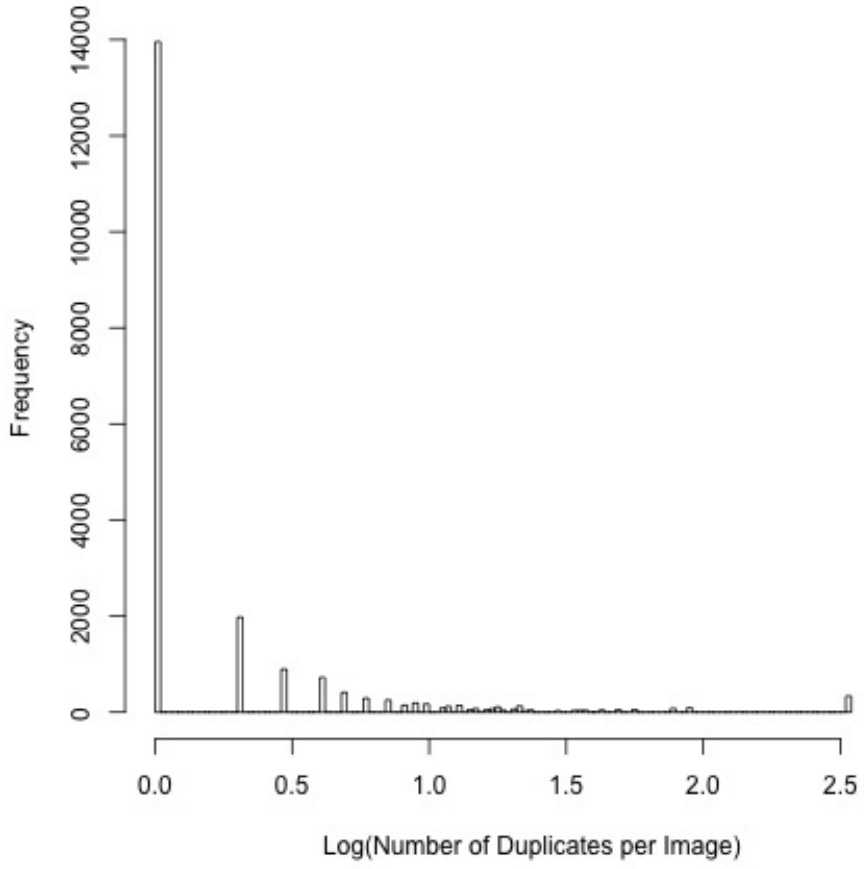
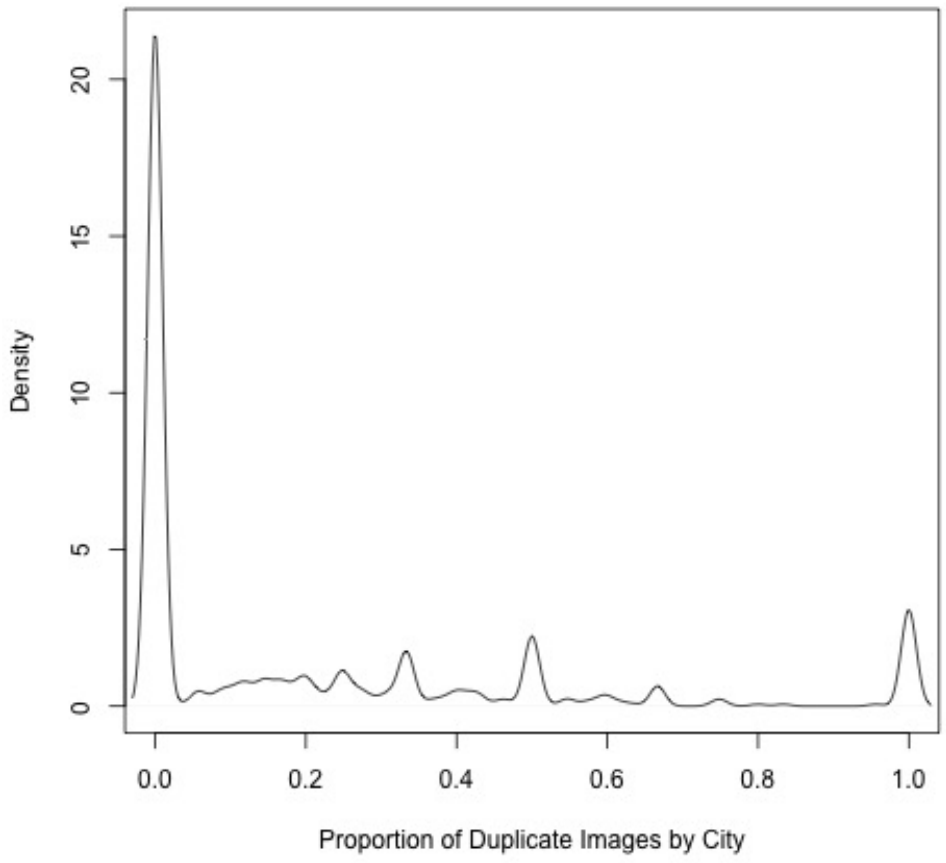


Fig. A17. Percentage of Duplicates by City



S8.3 Different Dependent Variable

Table A13 repeats the main model using three different operationalizations of the dependent variable. The first, shown in Model 2, does not log-transform the number of protesters. The results for the violence and free-riding variables are the same, except the number of photos containing police is now statistically significant; its sign matches the original model. Race Diversity$_{i,t-1}$ no longer correlates with subsequent changes in protest size.

The third and fourth models measure the size of protest using the number of users who share a protest photo per city-day. The third uses the raw count, the fourth logged. This quantity is smaller than the number of protest photos per day because users often share multiple photos. Results for violence and racial diversity are the same when not taking the logarithm, though they lose significance when log-transformed. Large Group$_{i,t-1}$ switches signs and is significant in both models, while Large Group$^2_{i,t-1}$ supports the same inference in Model 3 but not Model 4. Of all robustness checks in the manuscript and supplementary materials, these two differ the most from the original model.

These results differ the most from the rest of the paper's for two reasons. Most importantly, they embody a different data generating process then the other operationalizations of protest size. They do so because counting individual images provides less information about the size of a protest than counting the faces in an image. It provides less information because the photos are equivalent to randomly sampling a protest space and the surrounding protesters, akin to the leading methodology of in-person protest size measurement (Schweingruber and McPhail, 1999). Other work has shown that counting the number of protest photos less accurately recovers true protest size than summing the number of faces in those photos (Sobolev et al., 2019). Second, there is much less variation in this measure than in the sum of faces. The maximum value is 158, third quartile 1; when restricted to days with protest photos, the third quartile is 3.

S8.4 Accounting for Autocorrelation

Table A14 shows that the inclusion of lagged dependent variables up to 15 days old do not change the results for the violence or free riding variables already significant in the original model

Table A13. Different Measures of Protest Size

	Original	No Log	Number Users	$Log_{10}(NumberUsers)$
	(1)	(2)	(3)	(4)
Perceived Prtstr. Violence$_{i,t-1}$	$-.1674^{**}$	-9.8923^{***}	-3.4252^{**}	$-.1724^{***}$
	(.0677)	(2.8933)	(1.7108)	(.0552)
Perceived Stt. Violence$_{i,t-1}$	1.2820^{***}	86.4146^{**}	24.2096^{*}	$.4030^{*}$
	(.3327)	(37.2147)	(13.1782)	(.2269)
Perceived Stt. Violence$^2_{i,t-1}$	-2.1030^{***}	-168.2910^{**}	-45.4816^{*}	$-.6696^{*}$
	(.6093)	(79.2780)	(25.5827)	(.4007)
Police$_{i,t-1}$	$.7626^{*}$	125.1103^{*}	31.9542	.2401
	(.4493)	(74.2665)	(19.9307)	(.2776)
Fire$_{i,t-1}$	$.1009^{***}$	3.5316^{***}	1.2606^{*}	$.0690^{***}$
	(.0236)	(.5783)	(.7297)	(.0242)
Gender Diversity$_{i,t-1}$	$-.1126$	-9.8651	-2.0632^{***}	$-.0972^{***}$
	(.0939)	(7.3636)	(.7160)	(.0350)
Race Diversity$_{i,t-1}$	.0683	7.8310^{***}	1.0215^{*}	.0518
	(.0440)	(2.6819)	(.5586)	(.0348)
Age Diversity$_{i,t-1}$	.0203	1.6471	.6294	.0209
	(.0289)	(1.9407)	(.5215)	(.0212)
Tweets$_{i,t-1}$	$.0095^{***}$	$.2254^{*}$	.0112	$.0040^{**}$
	(.0033)	(.1340)	(.0432)	(.0021)
DV$_{i,t-1}$	$.1578^{**}$	$.0910^{***}$	$.2729^{***}$	$.3725^{***}$
	(.0682)	(.0348)	(.0310)	(.0973)
Intercept	$.1260^{***}$	3.5931^{***}	1.5215^{***}	$.1063^{***}$
	(.0237)	(.6761)	(.1740)	(.0106)
N	4,376	4,376	4,376	4,376
Adjusted R^2	.2450	.2051	.2741	.3710
City FE	Y	Y	Y	Y

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

(Model 1). (A partial autocorrelation plot suggested serial dependence for up to fifteen days.) The presence of police is now positively correlated with subsequent protest size. More noticeably, neither Race Diversity$_{i,t-1}$ nor Any Child$_{i,t-1}$ remain significant. This updated result supports the interpretation provided in the main paper of the cleavage variables: they are endogenous to the protests themselves, so controlling for enough previous protests removes those variables' significance.

Table A14. Robust to Additional Lagged Dependent Variables

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$	
	Original Model	15 Lags
	(1)	(2)
Perceived Prtstr. Violence$_{i,t-1}$	$-.1674^{**}$	$-.0750$
	(.0677)	(.0788)
Perceived Stt. Violence$_{i,t-1}$	1.2820^{***}	$.7844^{**}$
	(.3327)	(.3186)
Perceived Stt. Violence$^2_{i,t-1}$	-2.1030^{***}	-1.3370^{**}
	(.6093)	(.5303)
Police$_{i,t-1}$	$.7626^{*}$	$.6663^{*}$
	(.4493)	(.3987)
Fire$_{i,t-1}$	$.1009^{***}$	.0365
	(.0236)	(.0332)
Gender Diversity$_{i,t-1}$	$-.1126^{*}$	$-.0592$
	(.0939)	(.0616)
Race Diversity$_{i,t-1}$	.0683	.0279
	(.0440)	(.0430)
Age Diversity$_{i,t-1}$	.0203	.0130
	(.0289)	(.0338)
Tweets$_{i,t-1}$	$.0095^{***}$	$.0062^{***}$
	(.0033)	(.0022)
DV$_{i,t-1}$	$.1578^{***}$	.0588
	(.0682)	(.0360)
Intercept	$.1260^{***}$	$.0314^{**}$
	(.0237)	(.0129)
N	4,376	4,033
Adjusted R^2	.2450	.3220
15 Lags of DV	N	Y
City FE	Y	Y
Country Fe	N	N

*p < .1; **p < .05; ***p < .01

S8.5 Accounting for Days with No Protests

Table A15 shows attempts to account for days with no protest. Model 2 drops all days with no protest images. Model 3 is a Poisson model. Model 4 is a negative binomial, and Model 5 is a zero-inflated negative binomial model. To converge, Model 5 excludes city fixed effects and clustered standard errors; it does use country fixed effects.

Table A15. Count Models

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$		DV: $Sum\ of\ Faces_{i,t}$		
	Original	No Zero Days	Poisson	Negative Binomial	Zero-inflated Negative Binomial
	(1)	(2)	(3)	(4)	(5)
Perceived Prtstr. Violence$_{i,t-1}$	−0.167**	−0.209*	−0.798***	0.082	−0.775
	(.0677)	(0.121)	(0.120)	(0.641)	(0.504)
Perceived Stt. Violence$_{i,t-1}$	1.282***	1.175***	9.587***	10.271***	7.575***
	(.3327)	(0.447)	(0.343)	(2.569)	(2.089)
Perceived Stt. Violence$^2_{i,t-1}$	−2.103***	−1.834**	−17.328***	−15.779***	−12.593***
	(.6093)	(0.768)	(0.811)	(4.462)	(3.629)
Police$_{i,t-1}$	0.763*	1.519***	2.631***	1.843	2.121*
	(.4493)	(0.345)	(0.057)	(1.799)	(1.122)
Fire$_{i,t-1}$	0.101***	0.065**	0.124***	0.404**	0.272**
	(.0236)	(0.025)	(0.011)	(0.181)	(0.118)
Gender Diversity$_{i,t-1}$	−0.113	−0.135*	−1.315***	−0.285	−0.792***
	(.0939)	(0.078)	(0.045)	(0.479)	(0.300)
Race Diversity$_{i,t-1}$	0.068	0.073	0.513***	0.168	0.684***
	(.0440)	(0.050)	(0.027)	(0.321)	(0.206)
Age Diversity$_{i,t-1}$	0.020	0.025	0.666***	0.345	0.094
	(.0289)	(0.045)	(0.024)	(0.244)	(0.170)
Tweets$_{i,t-1}$	0.009***	0.006***	0.009***		0.012
	(.0033)	(0.001)	(0.001)		(0.010)
DV$_{i,t-1}$	0.158***	0.126***	0.002***	0.017***	0.002
	(.0682)	(0.046)	(0.0003)	(0.004)	(0.003)
Intercept	0.126***	0.507***	1.221***	1.018***	2.150***
	(.0237)	(0.045)	(0.024)	(0.181)	(0.117)
N	4,376	1,442	4,376	4,376	4,376
City FE	Y	Y	Y	Y	N
Cluster SE	Y	Y	Y	Y	N
Adjusted R^2	0.245	0.199			
Log Likelihood			−19,104.240	−4,387.215	−4,220.237

*p < .1; **p < .05; ***p < .01

S8.6 Weighted Results

Weighting each city-day observation by the number of protest photos shared from it strengthens the paper's results. Race and gender support critical mass theory. The free riding dynamics are more pronounced. The violence coefficients are much larger than the unweighted models, and model fit is almost 50% better than the paper's main models.

S8.7 Most Likely Protest Tweets

The last two models select tweets based on features that increase the probability they come from a protest. Model 5 restricts tweets to those only from mobile devices, based on the source

Table A16. Results Weighted by Protest Tweets per City

	.outcome			
	Original	Violence	Demographics	Combined
	(1)	(2)	(3)	(4)
Perceived Prtstr. Violence$_{i,t-1}$	−.1674**	−.3383**		−.3214**
	(.0677)	(.1379)		(.1357)
Perceived Stt. Violence$_{i,t-1}$	1.2820***	2.1774***		2.4331***
	(.3327)	(.3740)		(.3725)
Perceived Stt. Violence$^2_{i,t-1}$	−2.1030***	−3.9986***		−4.4522***
	(.6093)	(.6880)		(.6827)
Police$_{i,t-1}$	.7626*	1.3091***		1.2754***
	(.4493)	(.1440)		(.1438)
Fire$_{i,t-1}$	.1009***	.0198***		.0108***
	(.0236)	(.0140)		(.0139)
Gender Diversity$_{i,t-1}$	−.1126		−.2087***	−.2538***
	(.0939)		(.0517)	(.0499)
Race Diversity$_{i,t-1}$	.0683		.1434***	.1289***
	(.0440)		(.0326)	(.0318)
Age Diversity$_{i,t-1}$	.0203		−.0626**	−.0388
	(.0289)		(.0311)	(.0298)
Tweets$_{i,t-1}$	.0095***	.0026***	.0027***	.0030***
	(.0033)	(.0004)	(.0004)	(.0004)
DV$_{i,t-1}$	.1578**	.2564***	.2986***	.2791***
	(.0682)	(.0276)	(.0332)	(.0328)
Intercept	.1260***	.3940***	.5721***	.4906***
	(.0237)	(.0474)	(.0494)	(.0490)
N	4,376	4,376	4,376	4,376
City FE	Y	Y	Y	Y
Cluster SE	Y	Y	Y	Y
Adjusted R^2	.2450	.5533	.5253	.5675

*p < .1; **p < .05; ***p < .01

field Twitter provides with each tweet. If that field contains “Twitter Web Client" or “Hootsuite", the tweet is discarded; this paring leaves 3,743 tweets and 3,129 city days. The results for state violence and large groups match the full model, though with less statistical significance; the other covariates of interest lose statistical significance. The mobile model also fits the data less than half as well as the full model. Finally, we keep only tweets issued between 10 a.m. and 10 p.m., the most likely protest windows. Model 6 shows the results from these 4,664 tweets and 3,134 city-days. The result is a mixture of Models 4 and 5: the violence variables are larger and more precisely estimated, but none of the social cleavage variables are statistically significant, and the results for free-riding do not change.

Table A17. Most Likely Protest Tweets

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$		
	Original	Source Mobile	Protest Time
	(1)	(2)	(3)
Perceived Prtstr. Violence$_{i,t-1}$	−.1674**	−.0438	−.2242***
	(.0677)	(.1005)	(.0586)
Perceived Stt. Violence$_{i,t-1}$	1.2820**	.6405**	1.3883***
	(.3327)	(.3113)	(.4637)
Perceived Stt. Violence$^2_{i,t-1}$	−2.1030***	−1.0014**	−2.1913***
	(.6093)	(.4587)	(.8201)
Police$_{i,t-1}$	.7626*	−.0026	.6971*
	(.4493)	(.1366)	(.4005)
Fire$_{i,t-1}$	.1009***	.0578***	.1162***
	(.0236)	(.0214)	(.0278)
Gender Diversity$_{i,t-1}$	−.1126	.0168	−.0933
	(.0939)	(.0466)	(.0800)
Race Diversity$_{i,t-1}$	.0683	−.0126	.1058*
	(.0440)	(.0338)	(.0603)
Age Diversity$_{i,t-1}$	.0203	−.0066	.0088
	(.0289)	(.0246)	(.0285)
Tweets$_{i,t-1}$	.0095***	.0097***	.0100***
	(.0033)	(.0015)	(.0033)
DV$_{i,t-1}$	.1578**	.1121**	.1078
	(.0682)	(.0534)	(.0661)
Intercept	.1260***	.1374***	.1381***
	(.0237)	(.0144)	(.0161)
N	3,164	3,063	3,067
City FE	Y	Y	Y
Cluster SE	Y	Y	Y
Adjusted R^2	.2450	.1091	.2101

*p < .1; **p < .05; ***p < .01

S8.8 Models by Country

Pakistan is not included because it has too few observations.

S8.9 Investigating Fire, Police Variables

Table A18. Tables by Country

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$				
	Spain	Hong Kong	South Korea 2014-2015	Venezuela 2017	Venezuela
	(1)	(2)	(3)	(4)	(5)
Perceived Prtstr. Violence$_{i,t-1}$	.3920**	−.4626*	−.6093*	.1124	−.0955
	(.1941)	(.2711)	(.3254)	(.1231)	(.0681)
Perceived Stt. Violence$_{i,t-1}$	.7175**	3.7867***	−1.3734	.7740	.6684*
	(.3446)	(1.4235)	(1.4230)	(.5308)	(.3534)
Perceived Stt. Violence$^2_{i,t-1}$	−1.6441***	−8.2571**	2.6850	−1.1302	−.9714
	(.5217)	(4.1648)	(3.3816)	(.8877)	(.6119)
Police$_{i,t-1}$	1.2854***			.0876	−.1225
	(.1970)			(.4707)	(.2208)
Fire$_{i,t-1}$	.0178	.0715	.3637**	−.0451*	.0355
	(.0496)	(.0948)	(.1457)	(.0242)	(.0317)
Gender Diversity$_{i,t-1}$	.0642	−.0201	−.0101	−.1568*	.0172
	(.0777)	(.1480)	(.1195)	(.0866)	(.0855)
Race Diversity$_{i,t-1}$	−.0781	.1521	−.1716	−.0885	.0340
	(.0607)	(.1221)	(.1198)	(.0612)	(.0600)
Age Diversity$_{i,t-1}$	.0032	−.1156	.0316	.1542***	−.0456
	(.0372)	(.0908)	(.0665)	(.0525)	(.0490)
Tweets$_{i,t-1}$	.0047***	.0098	−.0003	.0153***	.0171***
	(.0011)	(.0072)	(.0035)	(.0033)	(.0038)
DV$_{i,t-1}$	.0680	.0763	−.0795	.0771	.0190
	(.0430)	(.1200)	(.0751)	(.0583)	(.0353)
Intercept	.5900***	.1034***	.0004	.0322	.0164
	(.0369)	(.0261)	(.0333)	(.0357)	(.0168)
N	1,412	257	365	573	1,752
Adjusted R^2	.2861	.1083	.1185	.6110	.0962
City FE	Y	Y	Y	Y	Y
Cluset SE	N	N	N	N	N

*p < .1; **p < .05; ***p < .01

Table A19. Fire and Police Variables on Their Own

	DV: $Log_{10}(Sum\ of\ Faces)_{i,t}$			
	(1)	(2)	(3)	(4)
$\text{Police}_{i,t-1}$	.4894	.4932		
	(.5121)	(.5071)		
$\text{Fire}_{i,t-1}$	.1011***	.0945***		
	(.0184)	(.0206)		
$\text{Gender Diversity}_{i,t-1}$		−.1148		−.1358
		(.0875)		(.0978)
$\text{Race Diversity}_{i,t-1}$		.0708*		.0968**
		(.0410)		(.0449)
$\text{Age Diversity}_{i,t-1}$		.0217		.0208
		(.0288)		(.0315)
$\text{Perceived Prtstr. Violence}_{i,t-1}$			.0237	.0227
			(.0627)	(.0649)
$\text{Perceived Stt. Violence}_{i,t-1}$			1.0602***	1.0529***
			(.3149)	(.3342)
$\text{Perceived Stt. Violence}^2_{i,t-1}$			−1.6560***	−1.6396***
			(.5213)	(.5458)
$\text{Tweets}_{i,t-1}$	.0095	.0096***	.0107***	.0108***
	(.0035)	(.0034)	(.0040)	(.0037)
$\text{DV}_{i,t-1}$	.1926	.1728**	.1954**	.1684**
	(.0743)	(.0704)	(.0805)	(.0711)
Intercept	.1239	.1275	.1224	.1273
	(.0158)	(.0158)	(.0159)	(.0159)
N	4,376	4,376	4,376	4,376
Adjusted R^2	.2373	.2389	.2304	.2334

*p < .1; **p < .05; ***p < .01

APPENDIX S9. INTER-CODER RELIABILITY

We used Fleiss' Kappa to measure the inter-coder reliability of our training image annotations. In many cases, the inter-coder reliability is typically measured on the coding data on which the actual analysis is conducted. In our study, the manual coding was performed on the training data, and the reliability was measured for the annotations to ensure that the models are trained in a consistent manner. Table A20 shows the estimated reliability statistics.

Table A20. Inter-coder reliability

Label	Kappa
Perceived Violence	.316
Perceived Protester Violence	.566
Perceived State Violence	.473
Large Group	.434
Small Group	.388
Police	.564
Fire	.702
Child	.457