Big Data is Watching You; Assessing the Dangers and Responses to Predictive Repression in Autocracies

Yousuf Abdelfatah
Georgetown University
Word count: 9000

March 18, 2021

Abstract

Repressive states are experimenting with machine learning to identify citizens with grievances who might join dissident movements. While they are already implemented in China, the viability of existing classification techniques and citizen data remain unassesed. We train classification and regression algorithms on trust in government and democracy preferences in 6 Arab countries using Arab Barometer data. Random Forest classification does outperform logit analysis, but the gains are moderate. Results suggest that the legitimacy costs of false positives will limit applicatiosn for the near future. Accuracy and cross validation suggests that detailed personal information about movement or social media use are necessary for effective applications.

1 Introduction Version 3

On June 25, 2017, a Xinjiang regional official of the Chinese Communist Party (CCP) circulated a bulletin about dissidents apprehended and sent to reeducation camps. While arbitrary detention in repressive states is common, these "dissidents" are the first selected for interrogation by an algorithm, rather than direct human identification. Officials were instructed to "put measures in place according to classifications (...) different types of tags pushed out by the "integrated" platform " (China Cables). On the basis of these tags and subsequent interrogation, tens of thousands were sent to reeducation camps in just that month.

Predictive repression in Xinjiang relies on "big data" collected from such diverse sources as police interrogations, internet use, searches of personal computers, government records, tax data, government services, utilities, biometrics and even the tracking of individuals locations. Scholars and journalists are now raising the alarm about a new pattern of predictive repression using "dataveilance" to preemptively repress actors. This new pattern diverges from the more common human analysis of speech and action for repressive targeting (Qiang, 2019). Students of repression speculate that preventative repression is now spreading from Xinjiang across China and to other aurocracies (Feldstein; Frantz et al.; Qiang; Wired). We refer to the automatic detection of potential future dissidents as predictive Repression.

The discussion of AR lacks a clear picture of which behavioral variables are most dangerous in the hands of repressive actors. Both scholarly and media descriptions of AR emphasize the number and diversity of behavioral variables that states can survey. However some variables the CCP currently uses lack a plausible connection

to grievances against the state (Wang, 2019). It is possible that new machine learning techniques change the predictive viability of big data. Nonetheless, we cannot assume that all used data is predictive because repressive actors can falsely overstate the effectiveness of AR to bluff complete knowledge of grievances.

A general idea of which data types are most important enables both researchers and citizens to better evaluate the state's claims. It also enables pro-democracy actors can influence dataveilance of certain states, through producing counter-surveillance software (Bock et al. 2019) and through state capacity development programs. A better understanding of variable importance allows us to target those efforts.

We directly measure the predictive value of a variety of citizen level variable for imputing grievances against the state. We train Random Forest models on survey data form 6 autocracies (Algeria, Armenia, Egypt, Jordan, Kuwait and Morocco). We left as x-data all survey questions which repressive actors could plausibly access, from internet usage habits to household location to charitable contributions. We use cross validation to assess the predictive value of the variables both separately and in general categories such as internet habits, movement, tax data, and service consumption.

There are serious limitations to our methodology. Most importantly, the dataveillance capacity of repressive states varies widely (Frantz et al. 2020). China has hired hundreds of thousands of security specialists to read internet communication and handcode data from checkpoints (Qiang et al, 2019; Greitens, 2020). In Egypt the state struggles to collect tax and employment data on a majority of citizens (Zaher, 2020). Our data best approximates an intermediate state between the two.

Secondly, survey data is not a perfect proxy for state dataveillance. Given the lack of alternative methods, we believe our technique gives the best possible measure of variable importance.

The main finding is that variables which describe either political behavior or information sources are more valuable. Tax and employment information had very low predictive value except for income measures. These results suggest that successful AR will depend on internet surveillance, not the direct collection of state data. Furthermore we show that when encrypted applications are widely used for non-political communication, detecting such apps does not predict grievances. Therefore the highest leverage area for preventing AR is improving access to digital privacy in repressive states.

The paper proceeds as follows. The next section explains why leaders impute the preferences of their followers. The following section briefly describes the history of PAR in Xinjiang, the only known use case. Next, we describe our data sources and analysis method. We then summarize the results, focusing on general trends in all 6 country cases. We conclude with comment on the CCP's motivation for using AR in Xinjiang and with policy priorities for preventing AR.

We define predictive repression as the use of machine learning to either predict future political behavior or impute beliefs. Predictive repression may use supervized or unsupervized learning. Predictive repression is a small corner of the broader digital repression category. In particular, predictive repression does not include punishing or censoring past public speech acts, a much more common tactic. New AI techniques are rapidly decreasing the costs of censorship and speech-policing. It should be

studied separately because it relies on distinct AI techniques and is intended to deter speech, unlike predictive repression.

Why Leaders Impute Beliefs

Predictive repression automates an ancient practice of political leaders: imputing the beliefs and preferences of followers. Actors prefer coalition partners who share their beliefs and preferences (Kahan 2013; Hanson and Simmler, 2018). Cobelievers are more loyal partners than nonbelievers, even for empirically viable beliefs such as climate change, policy impacts or the affects of political institutions. Untrue believes can even represent a *stronger* signal of loyalty because a non-member would not endorse falsehoods. We use belief and preference interchangeably because believing a regime is beneficial and preferring that regime are not distinguished in our argument.

Even in democracies, discrimination by belief can be strong. A 2010 survey found that 30 percent of Republicans and 24 percent of democrats prefer their children not marry across parties (Pew, 2014). Iyengar and Westwood find that American partisans are 60 percentage points more likely to award a scholarship to a copartisan (partisan bias was much lower in the past however). Timur Kuran argues that in autocracies norms against dissent are enforced socially, not just by repression specialists (although the state also discriminates) (1995). People highly prize membership in coalitions (identity groups, ideological camps, political parties and factions) in all polities as it offers security, connection, and other rewards. People thus have strong incentives hold socially-approved beliefs, especially in autocracies. On surveys, "political intimidation in China is real, of course, but its more likely effect is not to cause people to lie about what they think, but rather to shape the thoughts they have"

(Nathan, 2020).

People respond by both adopting utility-maximizing (safe) beliefs and hiding their true beliefs. Socially adaptive belief theory (SAB) refers to the strategic adoption of valuable beliefs. In SAB the actor sacrifices epsitemically accurate beliefs. In preference falsification, they disguise their preferences. Falsifying preferences retains an accurate underlying map of the world, but is dangerous because humans possess mind reading powers (through social cues, hesitation, speech patterns) (Williams, 2020). SAB is most effective when accurate beliefs are not valuable to the individual, such as in voting, revolution and punditry, but consequences for disloyalty are high. Coalition politics thus creates an arms race between discriminating coalition formers and potential members. Each actor wishes to know the true beliefs and preferences of others, while others prevent disagreement from surfacing (Kuran, 1995). Citizens do also employ direct deceit. Robinson and Tannenberg find that in a list experiment (where citizens have greater anonymity) regime support in China drops by 25 percentage points relative to a direct question.

Once grievances are imputed, autocrats can choose from a limited set of responses. Leaders often respond with repression to challenges to the status quo (Davenport, 2007), but a belief is not a challenge unless it is common knowledge. If grievances are imputed from non-political behavior, direct repression will disincentives such behaviors as association and phone use before personal views. Preventative repression should be particularly difficult to legitimate. The coerced confessions common in Stalin and Saddam Husseins regimes were plausibly intended to legitimate preventative attacks on suspected future challengers or dissidents. Furthermore, the regime

pays a cost when revealing grievances because dissidents become aware for their numbers and allies and have less to lose from stating their preferences (Kuran, 1995). The open question of whether and to what extent private grievance information benefits a repressive regime is beyond the scope of this paper.

An authoritarian regime can secure itself by placing loyalists positions of power, while marginalizing the disloyal. They also prefer to increase collective action capacity of loyalists and decrease that of dissenters. However, this goal is much easier stated than achieved. Evolutionary psychologists argue that humans possess innate evolved skills at coalition politics because it was critical to accessing safety and resources in early human tribes (De Waal, 2007). More recently, the collapse of the Soviet Union was proceeded in 1985 by a large shift in public opinion against communism (recorded in secret East German party surveys). The East German people successfully hid their beliefs until a sudden cascade caught the world by surprise in 1989. If the regime empowers many persons with socially adaptive beliefs or false preferences, it is in danger of a cascade. As a few persons of strong conviction enter opposition, the social rewards for loyalty decline (Kuran, 1995). The reduced loyalty incentive expands the opposition, creating a self-enforcing cycle that expands the opposition. This is true for preference falsifiers, socially adaptive believers and even true loyalists willing to preference falsify to an overwhelming opposition. Thus the regime desires the most stable underlying beliefs, if imputable.

Digital Repression

The digital communications world has opened a variety of tactics for repressive actors to dominate citizens by controlling information, punishing dissent and

degrading collective action (Frants et al.). Regimes commonly monitor digital communications, often by hand, to identify opposition leaders, activists and dissenters to target. Gohdes has already found that in the Syrian Civil War the Assad regime used targeted killings more in areas with higher internet access, and indiscriminate killings in areas with lower access. Activists and journalists in Egypt, Iran and elsewhere are routinely jailed for online statements. States are already monitoring digital communications to identify forming protests and proactively respond. Singapore is pioneering the use of facial recognition systems to identify protesters (Tan, 2020) and the Yanukovych regime in Ukraine mass texted all cell phones near protests in Kiev as an implicit threat (the threat failed to dissuade protesters). For a longer treatment of new tactics see Felstein (2019) or Frantz et al. (2020).

Franz et al. catalogued mechanisms of digital repression: identifying likely regime opponents by combining mass surveillance with machine learning; monitoring regime insiders; automating censorship, e.g. the Great Firewall of China; gauging public sentiment to anticipate and prevent protests; proactively spreading misinformation to disrupt collective action. Each mechanism is distinct and deserves a detailed assessment of its viability and likely affects. We know little about the effectiveness of digital repression. Most existing work categorizes and tracks the spread of new emerging technologies. Franz et al. constructed a global dataset of digital repressive capacity, but the technology is too new for a statistically identifiable affect on macropolitical outcomes. Predictive repression is just one mechanism of digital repression and is relatively new and untried.

The predictive imputation of ideology diverges from existing strategies. First, the

intended targets are all members of society, not the first movers of a revolution who tend to be highly ideological and risk-tolerant (Kuran, 1991). Secondly, the regime does not wait to observe revolutionary acts or statements then punish them. Instead it proactively alters beliefs and capacity by marginalizing persons with a particular ideology, in this case through mass arbitrary detention.

Imputing beliefs throughout a society presents deep practical challenges. Autocratic politics should select for leaders skilled at imputing the loyalty of potential followers, as this helps the would-be autocrat rise up. But even a skilled leader can only assess a few hundred close followers. Down the hierarchy she must create a principal agent chain of loyalty assessors. The Soviet Commissars are am example specialist institution for policing loyalty. Modern autocracies use secret policy ministries, sometimes several (sultanistic regimes guy) or a single political party to recruit and indoctrinate member-informants (Geddes). Secret policy are expensive and many autocracies struggle to form strong parties despite their effectiveness. A primary role of both institutions is to extend the imputation of loyal beliefs down through society and implement discrimination.

Loyalty imputation has already changed with the move to digital repression. We show that while many regimes take advantage of digital communication, much of the imputation is done by humans.

Predictive repression should increase regime durability if it:

- Is cheaper to implement than traditional belief imputation
- Can be applied more widely, accessing previously unmonitored social strata

- Is more accurate than traditional methods (Jung at al. 2017) find that regression-derived checklists outperform intuition in bail setting)
- Makes belief-based discrimination more acceptable to the public.

In both competitive and non-competitive polities, citizens value government fairness highly (Nathan, 2010). The last Asia Barometer round found that "in nine of the fourteen countries, citizens give even more weight to government fairness than they do to economic performance" (Nathan, 2020; 162). Punishment or discrimination based on hunches about unstated beliefs violate widely-held justice norms. In elite politics coerced confessions or corruption allegations evade those norms. The legitimacy costs of AR-based discrimination in the general population are untested but expected to be large.

¡where does this fit¿

Predictive ideology imputation is just one new tool within digital repression (Feldstein, 2019). Digital repression is "the use of new technologies, primarily the Internet, social media and Artificial Intelligence (AI) - to (...) maintain political control. While the techniques are new, the goals are constant: to raise the costs of disloyalty, to identify the opposition, and to prevent collective action against the regime. The costs of repression include the risk of backlash as unpopular punishments and surveillance enlarge the political opposition. Traditional repression also requires the regime pay, train and motivate thousands of party members or spies to surveill citizens. Furthermore, repression prevents citizen collective action, such as forming companies, universities and other associations. That atomization of society has negative consequences for growth (Acemoglu). Digital tactics can improve repression by reducing

any of these costs.

1.1 Predictive Repression Today

This section summarizes public knowledge about AR in Xinjiang, the only verified implementation. ¡summarize findings below¿

There is a long history of repressive escalation, deescalation and dissent in the resource-rich Muslim-majority province of Xinjiang (Greitens et al. 2020). The most recent episode of major public resistance occurred in 2008-2009 with attacks on police stations and violent clashes between Uighur and Han. An estimated 200 people were killed in the police response. In response to this mass resistance the CCP rapidly increased local coercive capacity using its customary techniques, such as increased security spending and stationing a trained policemen of local origin in each village. Mass imprisonment and intensive personal data collection were not part of the response. Tibet experienced the same timeline of resistance and response.

Although public contention declined after 2009, Uyghur participation in Islamic terrorism increased. Several high profile terror attacks occurred within China and some 300 Uighur travelled to Syria to fight with ISIS. Greitens et al use CCP public officials discourse to argue "Around 2015-16, just as the CCP observed new evidence of Uyghur participation in Islamic militant groups abroad, it also concluded that as much as a third of Xinjiang's population was vulnerable to extremist influence (...) Officials concluded that existing policies focused on degrading citizens' capacity for terrorism were inadequate" (pp. 44).

In 2016 repressive tactics in Xinjiang diverged from Tibet and the rest of China.

Over one million Uighur were involuntarily detained in re-education camps where they are subject to political indoctrination. The state tolerates a high false positive rate in identifying targets for detention, on the basis of eradicating Islamist ideology in China. The primary role of the AR tactics described below is selecting targets for detention and reeducation. Therefore the primary purpose is to identify a particular anti-state ideology. The system then pushes out "tags" on certain individuals who are then arrested and moved to reeducation camps.

We possess detailed knowledge of mass surveillance in Xinjiang from a mobile app used by security forces which Human Rights Watch accessed and reverse engineered (Wang, 2019). In addition a small number of internal party documents have been leaked (Wilson-Chapman, 2018). The app, named the Integrated Joint Operations Platform (IJOP), collects data on a surprisingly wide set of personal activity. Some entries are plausibly connected to subversive behavior such as sharing "Wahhabism", "knowing how to make explosives", and possessing foreign messaging applications. But many entries lack plausible relevance to regime support: "unwilling to enjoy policies that benefit the people", "Collected money or materials for mosques with enthusiasm" and "household uses an abnormal amount of electricity". Data is often collected during intrusive home visits which families cannot refuse. Much of this data is collected by hand in a laborious process of checkpoints and home visits by tens of thousands of contract security workers (Wang, 2019). Security personnel also routinely search residents phones for suspicious applications.

The IJOP also relies on automated data streams from public and private service providers (Wang, 2019). It imports data from; CCTV cameras equip ed with fa-

cial recognition software; visitors to residential areas and schools; police checkpoints; package shipping; detailed information about package shipping; electricity consumption and gas station visits. When an "unusual" amount of electricity use is detected officers are dispatched to investigate the household and seek a plausible explanation. Biometric information including DNA samples, fingerprints, iris scans, blood types and voice samples is also collected, but lacks a plausible predictive value.

We know much less about how the IJOP analyzes this data. The app itself contains only simple conditional statements for investigation tags, such as "if the person who drives the car is not the same as the person to whom the car is registered, then investigate this person". We also know that the implicit probability threshold for detention is very low. For example, reports suggest that any person possessing the messaging application WhatsApp is detained. However, the central IJOP system could use much more sophisticated algorithms to analyze this personal data. Leaked documents show that the IJOP releases tags for individuals who are then investigated and a large portion are detained for reeducation (Wilson-Chapman, 2019).

The majority of the monitored behaviors lack any obvious connection to political ideology, dissent or terrorist activities. One explanation is that the security forces are detaining persons almost indiscriminately, while falsely claiming to have a sophisticated detection system (Wang, 2019). That explanation fits Greiten et al.'s narrative that the party is happy to detain many innocents to eradicate Islamist ideology. An alternative explanation is that the CCP can use modern AI techniques to identify ideologies from apparently unrelated data.

This technology is likely to spread outside of Xinjiang if it provides real repressive

value. Xinjiang is a first training ground because China is an AI development hotspot and the CCP tolerates a high false positive rate in Xinjiang. As the technology spreads and/or specificity improves, other regimes will copy these tactics.

China is already exporting the technology to implement predictive repression elsewhere (Feldstein, 2019; Polyakova and Meserole, 2019). Singapore, Malaysia, Zimbabwe and Dubai are importing facial recognition systems from Chinese state contractors .Venezuela is contracting with Chinese firm ZTE to build a "national ID card, payment system, and "fatherland database" that will track individuals' transactions alongside personal information such as birthdays and social media accounts" (Polyakova and Meserole, 2019, pp. 6; Berwick, 2018). Ethiopian security services use use ZTE provided tech to digitally monitor opposition activists (HRW, 2014). Qiang suggests that China's social credit system, ostensibly for incentivizing prosocial behavior, could easily extend the IJOP to the rest of China (Qiang, 2019). While these applications have yet to begin automatically imputing grievances, they underline the threat of global spread.

Research design

The main objective of this study is to determine broadly which types of data are most concerning for predictive repression. We roughly simulate grievance imputation using publicly available data on observable behaviors and political attitudes in authoritarian states. We then extract variable importance information from the most predictive models for comparisons

We simulate grievance imputation using a "split, train, test" process. The input variables are divided between X-data, which is used for imputation, and Y-data which contains the true outcome information (in this case grievances). The observations (respondents) in each data set are randomly split between a larger training and smaller testing dataset. The predictive algorithms use both the X and Y data in the training set to construct their models. The model then uses the X data for the test set to impute Y values. We then analyze the accuracy of the model in predicting the Y-values.

Data

In democracies individuals have little incentive to hide their ideological positions and often intentionally signal their loyalty and belonging. In repressive states individuals have strong reasons to hide their ideology. Because ideology imputation relies on public behavior, tools for predicting ideology in democracies should be less effective in autocracy. For example, choice of newspaper should be less predictive when the state closes dissenting newspapers. As Kuran argued "In [repressive states] the very forces that discourage truthful expression also inhibit the collection and dissemination of opinion data" (1995 p. 1538). Therefore we only used data from countries with polity IV scores below 5 (Davenport and Armstrong 1996; Regan and Henderson, 2002).

We selected the Arab Barometer (AB) Wave V survey data from 2018. The AB covers a variety of states which rely on repression and limit mass political participation. It provides detailed information about political beliefs (for classifying

grievances), political and economic behavior and information consumption. It has especially detailed information about internet use. The survey questions vary only slightly between countries.

We required countries where a salient ideological divide exists between pro-democracy and anti-democracy positions. We expect ideologies to be more predictable when discussed or personally experienced, and respondents may give random answers to non-salient ideologies. The Arab countries share a salient ideological divide about the degree of democracy. In Sudan, the survey was completed 12 days before the 2018 revolution began with an 8-month mass civil disobedience campaign for democracy. In Algeria in 2019 millions entered the streets in a pro-democracy cascade (Volpi, 2020). Ideological positions toward democracy are also shaped by life experiences. In Egypt, Ketchley and El-Rayyes (2020) find that sustained protests for democracy reduced support for democracy by disrupting daily life.

We did not model democracies, defined by a Polity IV score above 5 (Tunisia and Lebanon). We also did not model countries experiencing civil wars (Iraq, Liby, Yemen) or foreign occupation (Palestine). This left 6 states, Algeria, Egypt, Jordan, Kuwait, Morocco and Sudan. After removing null responses, each country had at least 1400 responses suitable for use, and all but Kuwait had over 2000.

Y Variables

Our Y-variables were inspired by recent work on citizen grievances in repression (Gregory et al.; Rozenas et al; Aspinall). We interpret a grievance as a preference for anti-regime action or for regime change. Following conventions in formal theory, a grievance is the positive dummy variable (Rozenas). For robustness we included

two grievance metrics as y-variables.

The trust in government y variable is based on trust in the "Government (council of ministers)", as reported by AB. A grievance is recorded for the answers "not a lot of trust" or "no trust at all", otherwise no grievance. The regime preference variable is positive if the respondent preferred the statement "Democracy is always preferable to any other kind of government" over "Under some circumstances, a non-democratic government can be preferable" or "For people like me, it doesn't matter what kind of government we have". The respondent must also report that the country is below six on a ten-point democracy scale. In all AB countries very few respondents ranked democracy as a priority issue, and relative issue ranking was not included in the grievance variables.

The regime-preference variable has greater specificity, as it excludes low-trust citizens who still prefer the current regime. However, respondents who prefer a different nondemocratic regime are excluded from the regime-type preference specification and potentially included in trust specification. Because authoritarian rules face threats from both democratic and nondemocratic challengers, we included results for the trust specification.

X variables

We included in the X variables all behavioral information that could plausibly be collected by the state. Behaviors were included if a high-capacity autocracy such as China could plausibly gather the information. We did not vary the inclusion criteria by the capacity of the country of data collection. Wherever possible we used verified accoutns of state surveilance to decide inclusion criteria.

The variables user_facebook, user_youtube, user_whatsapp, user_twitter, user_snapchat, user_instagram and user_telegram identify users of said apps. App usage is a target of the IJOP surveillance system, where it appears that possession of an end-to-end encrypted app is sufficient evidence for arbitrary detention of Uighurs. The first author was once denied a visa to Israel after border police identified the encrypted messaging app signal on his phone. The worst instance of app surveillance is the 2020 Belarus protests where the dominant messaging app was not encrypted, forcing activists to use the secure telegram app. Unfortunately the police responded by accosting citizens in the street and searching their phones for telegram (Euractiv, 2020). In extreme cases citizens were tortured to access their phones. In all identified cases the regime either coerced citizens to give access or used a physical connection to hack into the phone, rather than surveiling app usage digitally en masse. The diversity and security of cellphones may inhibit mass surveillance for the moment.

Repressive states also spend significant resources monitoring social media (Frantz, et al.). One report leaked that the CCP employed 2 million persons to monitor digital communications identified by keyword in 2014 (Xu and Albert, 2014). Our variables socmed_use and internet_use measures numbers of hours of use per week, respectively. Internet is a binary variable for access to internet. Infs_socmed is true for all respondents who reported social media as their main source of news. We argue this course data gives a rough indication of the variable importance of social media data relative to other variables. Furthermore, current revolutionary organizing is quickly moving away from surveillable social media applications. The Belarus revolution of 2020 was organized mainly on Telegram. Iranians strongly prefer to use

Telegram for both texting and commenting, and Iranians have repeatedly defeated state attempts to block the app (Radio Free Europe). We revisit the surveillance of social media in our conclusion.

We include data on political behavior such as petitioning, protesting and voting. Unfortunately only local voting behavior was available in the Arab Barometer data set. Petition signatures are generally public knowledge. Voting behavior is often secret but could conceivably be suveilled, especially when the regime controls the voting format. We included protest attendance in light of a 2014 incident in Ukraine in which the regime mass texted all phones in the area of opposition protests with a threat. The Chinese government is also developing facial recognition software to identify protest participants (Frantz et al. 2019).

We also included a variety of data that states routinely collect in the process of tax and service provision. These include incmoe, education (educ), age, gender, chaitable contributions as a dummy variable, marriage, and employment. We further breakdown employment status between housewife, unemployed, student, retired, self employed and between the private and public sector. The variable orgmem records membership in an organization or club, which are targets of the IJOP surveilance system.

- Demographics: age, gender, religion, marital status, neighborhood income, residence in the capital province
- Tax information: income, retirement status, employment status (public, private, self or none), student status

- Social behavior: education level, charitable contribution
- Access to information: newspaper reading (Borzyskowski and Kuhn), social media use (hours per day), radio use, and television use, internet access
- Political behavior: petitioning, protesting, campaign rally attendance and voting in local elections

We excluded from x-data all responses based on the private beliefs or preferences of respondents. That includes trust in other institutions, beliefs about women's rights, and preferences for other state institutions. It is possible that NLP will grant access to some such preferences to future AR implementations using social media statements. However this data would be null for most person-variable dyads because social media statements are much fewer than all possible beliefs. Supposing that 1% of citizens tweet about their trust in capitalism, using those tweets would require an inverse increase in training data. Training data is constrained because unbiased information about citizen beliefs is expensive. Social media data which reveals grievances directly do not require prerdictive imputation and are beyond the scope of this study. Secondly, we would expect trust in other institutions to correlate with trust in regime if citizens do not distinguish within the state, regardless if regime-change preferences.

Certain questions were only usable in some countries. Some questions were not asked in all countries (local voting and Palestinian or Jordanian descent). Questions were also dropped if more than 100 respondents refused per country. Algerian respondents were not asked about trust in government. Kuwaiti respondents were not

asked about their preferred political system.

The survey data may suffer from preference falsification or self-censorship. Robinson and Tanenberg find self-censorship of 25% in list experiments in China. However, we would not expect self-censorship to systematically bias the predictive value of different variables, particularly across the five countries. Furthermore, any actual implementation of AR would suffer from similar or higher levels of data falsification. The Xinjiang leaks show that citizens evade surveillance by faking documents, discarding their phones, and registering under the identities of dead relatives (Wilson-Chapman, 2019).

Data Analysis

First we identified the most effective classifier algorithm for the dataset using cross validation. Cross validation allows us to estimate the best model before running the test data through it by "holding out" a subset of the training observations and then using them as a stand-in to test. The Cross-Validation method used here is k-Fold validation, which splits the observations into k groups, with k-1 groups used to train the model and the last group used as the test set. This is then repeated k times, with the cross validation estimate computed based on the average of the k test groups. For the purposes of this analysis, k was set to 5.

The classifiers that were tested were:

- Naive Bayes- a probabilistic classifier based on the Bayes theorem. It calculates the probability that there is backsliding, given the values of the independent variables, assuming that the independent variables are also independent from

one another

- K Nearest Neighbors- a classifier that identifies the number of 'K' points from the training data that are closest to the observation of interest, and then determines its category based on the majority category of the nearest points.

- Decision Trees- A classifier which sorts observations by splitting them based upon specific decision criteria. It then predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs

- Random Forests- A classifier which builds several different decision trees by pulling several training sets from the training data along with a random number of predictors. It then creates its decision criteria by averaging across the predictions from each tree

- Support Vector Machine- A classifier which sorts objects into categories by creating "decision boundaries" distinguishing between classes.

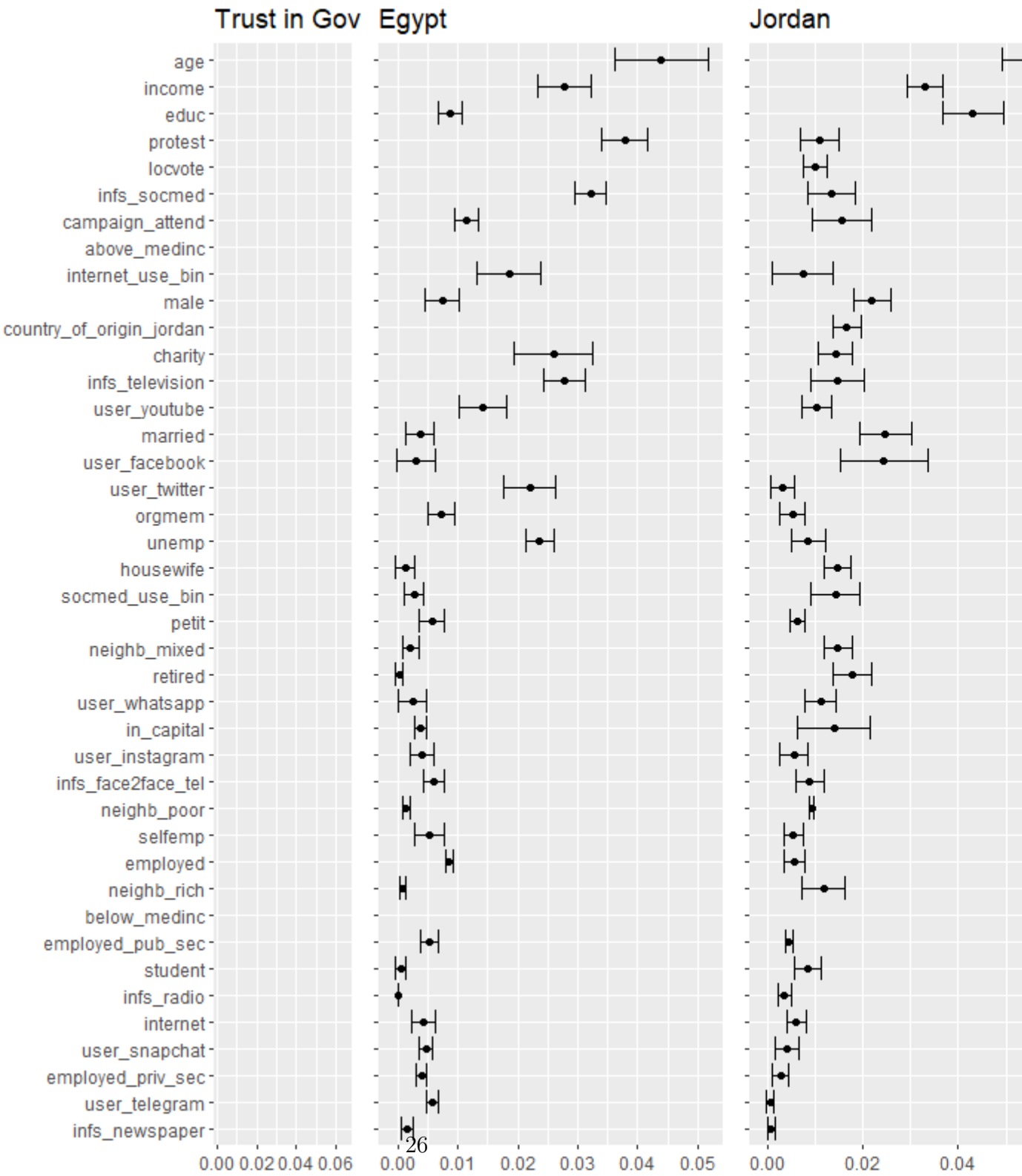
In order to evaluate the performance of each algorithm, the full data set was split into two subsets: a training set and a testing set. The algorithm was first fit on the training set, and then based on the information it gained from the training data it predicted the categories of the test set. The model was then judged on its ability to correctly the predict the dependent variables of the test set. The segregation of training data from testing prevents over fitting so model performance is accurately scored.

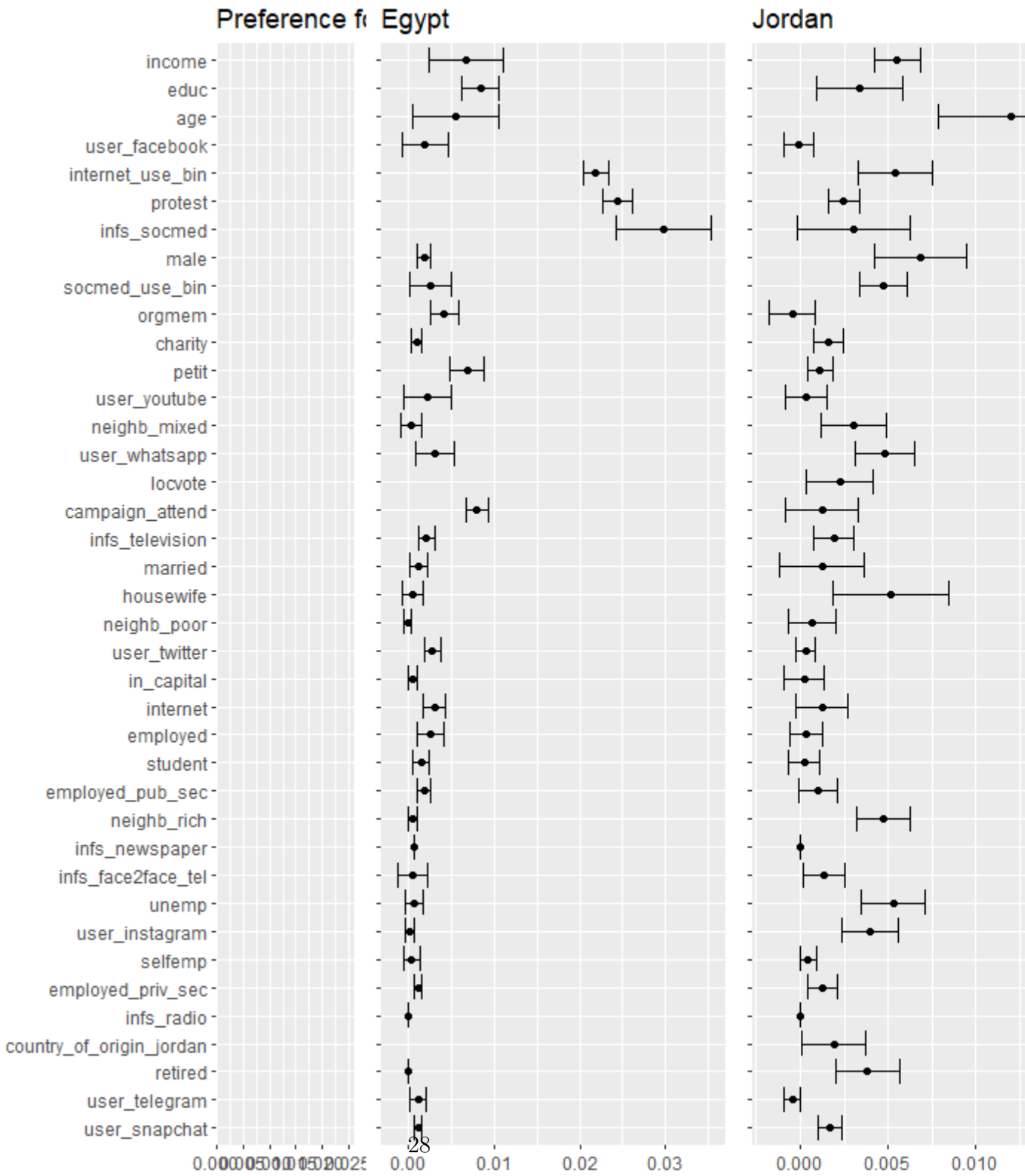
The Random Forest algorithm outperformed all other classifiers in every data set. The remainder of the paper focuses on them.

Decision trees consist of a series of splits on the original dataset, known as a tree. Each split, which can be thought of as a branch, is made on a "decision node." This decision node denotes a criteria that the data is being split on, for example countries with a GDP per capita greater than 25,000. In that countries with a GDP greater than 25,000 would go on one branch, while countries with GDP less than 25,000 would go on another branch. These groups would then be split on another criteria. The farthest branches of these trees are referred to as "terminal nodes" or leaves. In this way a decision tree is actually like an upside down tree, with the leaves on the bottom. The model utilizes a "top down" and "greedy" approach in deciding what features to split on. Put simply, this means that starting from the top of the tree, the algorithm makes each split based on what best minimizes classification errors at that specific step (ie grouping as many 1s together and 0s together as possible while minimizing the number of members of the other class in the group).

While Decision Trees have the advantage of being both relatively easy to interpret and a closer approximation of human decision making, they lack the predictive accuracy of other models, are sensitive to small changes in the data, and tend to overfit, especially as the tree gets deeper and decisions are made on smaller and smaller subsamples of the data. The Random Forest model improves on decision trees by building many trees, each created from a sub sample of the data, then merging them together and taking the most common predictions across each terminal node for each tree. Each tree is based on a random subset of the independent variables. The added

variation caused by training on different samples of the data and utilizing various subset of the variables allows the model to create a more accurate and stable prediction upon taking the average of all the trees. The use of random subsets which are later combined also helps prevent overfitting . Additionally, the algorithm was tuned on several parameters in order to maximize its predictive power. These parameters are the maximum depth of each tree, the number of trees to build, and maximum number of features to consider for each tree. The ideal parameters differed across countries and y variables.


Trust in Gov
Egypt
Jordan
age
income
educ
protest
locvote
infs_socmed
campaign_attend
above_medinc
internet_use_bin
male
country_of_origin_jordan
charity
infs_television
user_youtube
married
user_facebook
user_twitter
orgmem
unemp
housewife
socmed_use_bin
petit
neighb_mixed
retired
user_whatsapp
in_capital
user_instagram
infs_face2face_tel
neighb_poor
selfemp
employed
neighb_rich
below_medinc
employed_pub_sec
student
infs_radio
internet
user_snapchat
employed_priv_sec
user_telegram
infs_newspaper
0.00 0.02 0.04 0.06
0.00 0.01 0.02 0.03 0.04 0.05
0.00 0.02 0.04


Preference f
Egypt
Jordan
income
educ
age
user_facebook
internet_use_bin
protest
infs_socmed
male
socmed_use_bin
orgmem
charity
petit
user_youtube
neighb_mixed
user_whatsapp
locvote
campaign_attend
infs_television
married
housewife
neighb_poor
user_twitter
in_capital
internet
employed
student
employed_pub_sec
neighb_rich
infs_newspaper
infs_face2face_tel
unemp
user_instagram
selfemp
employed_priv_sec
infs_radio
country_of_origin_jordan
retired
user_telegram
user_snapchat
0.00
0.01
0.02
0.03
0.000
0.005
0.010

1.2 Trust in Government Results

We depict the permutation importance measures from the random forests for each y variable in figures 1 and 2. The test is described above. The variables are listed in order of their unweighted average importance.

The most striking result is the variance in variable importance between countries. Protest attendance is the second most revealing variable in Kuwait, but is within two standard deviations of the null in Sudan for trust. Twitter is use is predictive in Egypt, but near null in Jordan, Morocco and Kuwait. Charity and unemployment are highly relevant only in Egypt. Teasing out the reasons for this variation is a task for regression analysis. The variation limits our ability to generalize about legitimacy imputation.

Income, age and education have high importance in all countries. This is not surprising given how those variables strongly shape life experiences, economic interests, and political knowledge. Gender and charitable contributions also achieve consistent significance across the cases. Surprisingly, marriage and employment sector all perform relatively poorly. Although they are usually distinguishable from noise, the employment variables rarely pass .01 units of importance.

As expected, the political behavior variables perform well. Protest attendance was the third most important variable on average, and voting the fourth. Organization membership performed poorly, likely because the question is vague. Petitioning also preformed poorly, possibly because it is rare. The highest petitioning rate oc-

curred in Morocco at one in 5. Campaign attendance had varied performance, highest in Algeria. One explanation is that voting and campaigning in the authoritarian countries is usually more clientelistic than ideological (Blayeds, 2006).

The informational and internet variables perform moderately. Use of social media as information source is the 6th most important on average, and number of hours of internet use is 9th. In Kuwait social media news sources is second, and in Egypt third, but in the other three states it performs poorly. The various apps all have moderate or low performance.

In general, these results suggest that trust or attitude imputation is sensitive to diverse set of variables. This validates Qiang's concern about the variety of data sources being gathered by repressive actors (2019).

1.3 Regime-type preference results

The models predicting of regime-type preferences found very different results. The preference for democracy variable is 1 for all respondents who prefer democracy and rate their current government a 5 or less on a 10-point democracy scale. It therefore captures more of an ideological position than an attitude, and is positive for between a quarter and a third of respondents.

The non-political and non-informational variables perform more poorly at predicting ideology. While income, age and education remain the strongest predictors on average, their lead is greatly reduced. The average value for age dropped from .07 to .015, and income is reduced by half. Education has a more moderate reduction, perhaps because it strongly interacts with information consumption. The employment

and location variable have moved entirely to the bottom half of the distribution, and the highest financial variable other than income is charitable contributions.

Those variables relating to political behavior and information consumption perform much better. Simply the use of Facebook is now the fourth most important variable, and is most important in Morocco. Hours per week of internet use is now the fourth most important on average, followed by protesting and social media information sourcing. The starkest illustration is Egypt, where hours of internet use, protest attendance, and social media use are each 4 times the value of any other variable. None of the non-political and non-media variables pass .01 in Egypt.

Except for Facebook, the time and information sourcing on social media is more relevant than the particular app. Twitter has low relevance even in Egypt where 439 respondents use it. Telegram and Snapchat have low importance because they are rarely used in the particular countries. Despite being end-to-end encrypted at the time of surveying, whats-app is not highly predictive. We dio not reject the null in Algeria and Morocco and performance in Algeria and Sudan is modest. This demonstrates that once secure apps become popular across the population, it cannot identify opposition ideology even with nonlinear models. This observation is critical to countering state surveilance and we return to it in the conclusion.

These results suggest that non-political data has weak relevance to ideology in autocracies. Figure three summarises this result. One explanation is that most citizens of non-democracies do not form strong ideological positions on the basis of life experiences and material interests, because they lack safe avenues for political expression. Therefore while life experiences condition attitudes toward the state,

they do not affect ideology. Ideology does have a two-way relationship with news consumption and political expression. More ideologically minded individuals seek out news and news changes ideological positions.

Type of Behavior	Trust in Gov	Regime-Type Preference
Political behavior	High	High
Non-Political behavior	High	Low

This suggests the danger of regimes imputing opposition ideology from non-political data is low. Unfortunately data limitations prevent assessing the variety of behaviors states now seek to monitor from purchases to finances to shipping and geolocation. It remains possible that accumulating enough low-importance data could overcome the low value per observation. But consistent accuracy is less likely given our results.

1.4 Conclusion

¡summarize method nad results¿

1.5 Protecting Private Communication

The highest leverage intervention to prevent effective predictive repression is to protect the privacy of political communication. Our results suggest that monitoring expression and information consumption is more viable than nonpolitical behavior. Furthermore pro-democracy actors can effectively enable citizens to protect their communication. Unfortunately we have few options to prevent autocrats from analyzing financial data, public service use and police interviews.

Our results suggest that increasing the encryption of personal communication is an effective response to predictive repression. An end-to-end encrypted messaging app (Whatsapp) had only 14th highest average var imp for ideology, below YouTube and Facebook. Whatsapp was also used by more than 40% of respondents in each country except Algeria. This is surprising since both the Belarus and Chinese states hunt for encrypted app users. The explanation is that if secure communications are popular enough in society they cease to signal any political position. In Belarus activists had to move to Telegram during the revolution because the dominant app (Viber) is not secure. Incredibly, the people of Iran gamed this out beforehand and preemptively moved to telegram in 2015.

Therefore spreading secure applications has the double benefit of protecting the particular communication and camouflaging a privacy-hungry opposition. But if only opposition members use secure communication it becomes itself an opposition marker for regimes to exploit. The best case scenario is to ensure most digital communication is secure as the norm in each state, which is possible because communications apps have higher barriers to entry.

The good news is that secure communication is already the norm in most autocracies. Whatsapp, Telegram and Facebook messenger all offer encrypted communication (as an option in Facebook's case). ¡summarise the current situation¿

Our results show that the news readers choose is valuable information. Whether respondents received news mainly from social media was the 7th most predictive variable. Therefore pro-democracy actors should be concerned about information consumption, not just expression. States could monitoring news site visits, follows

and or censored TPS packets per internet user. Even if states lack training data, they could simply label opposition websites using the same apparatus as web censorship.

VPN subsidies would both protect encrypted apps and prevent information consumption monitoring. Iran has attempted to block Telegram, forcing users to either use virtual private networks (VPNs) or unsecured "forks" of Telegram. The attempt failed as Telegram retains a huge share of internet traffic (60 by some estimates). Nonetheless, such attacks remain a major threat to the privacy of ideology. A promising response is to temporarily subsidise VPN traffic in any country that launches a censorship attack on secure communications. Telegram CEO Pavel Durov has already piloted such a subsidy. A credible commitment to subsidise VPNs would be a strong deterrent to censoring encrypted data. If citizens respond to banning by adopting VPNs, the regime also loses the ability to monitor news. Furthermore, failing to ban the application would embarrass the regime and embolden the opposition. increased VPN use embarrasses the regime and prevents other forms of surveillance. If predictive repression becomes common, a general VPN subsidy for repressive states could be a cost-effective response.

1.6 Other Data Types

The charitable donations may predict ideology because people donate to opposition affiliated charities. The oppositions of many Arab countries organize large charity networks, like the Muslim Brotherhood in Egypt and Jordan. Those actors donate through cash o other difficult-to-monitor mechanism if predicitve repression monitors financial transaction as Venezuela hopes to. Training opposition movements to

operate outside of the surveilable financial sector could also be high leverage. Alternatively, the charity effect operates through personality correlations this intervention is unnecessary.

Geolocation data remains a serious concern. While the residence variables performed poorly, protest attendance was highly predictive. The Ukrainian incident already shows that regimes will use geolocation data (Walker, 2014), and Singapore has purchased Chinese facial recognition software to identify protesters.

States are also monitoring loans, purchases, shipping, service use and utility consumption. Our results weakly suggest that information with no apparent political relevance has weak prediction on ideology but some for legitimacy, We advise against a strategy focused on protecting such data currently. Pro-democracy actors have few avenues to prevent autocracies from using their own data or coercing data from service providers. Our best hope is to monitor machine learning articles coming from China.

1.7 Prediction Within Digital Repression

In the short term there is little evidence that supervised learning to predict political ideology is the most dangerous. So far it has only been implemented in Xinjiang for a mass incarceration campaign, where demands for precision were low. Meanwhile other digital repression strategies which automate the punishment of dissent, the detection of protesters or misinformation campaigns are already operating around the world. These tactics are already are already shoving efficacy and should be higher priorities than digital repression.