# Essays on Machine Learning in International Conflict and Social Networks

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Daniel Kent, M.A.

Graduate Program in Political Science

The Ohio State University

2020

Dissertation Committee:

Bear Braumoeller, Advisor

Skyler Cranmer

Christopher Gelpi

James Wilson

## Abstract

This dissertation leverages developments in machine learning methods to better model networked social processes, with an emphasis on international politics. The first chapter develops a dataset with estimates for every country's level of dissatisfaction with the international system from 1816-2012. The second chapter takes these dissatisfaction measures and uses them as features in a machine learning model which predicts international conflict onset. The third chapter explores spillover effects in social networks, demonstrating how causal forests can be employed to uncover spillover effect heterogeneity. Across these chapters, machine learning techniques are instrumental in modeling outcomes of interest and leveraging information from social networks.

In the first chapter, I propose a novel measure of international dissatisfaction spanning from 1816 to 2012 which explicitly operationalizes Gilpin's framework: the difference between a state's expected and actual benefits from the international status quo. I estimate a state's expected international benefits by building upon recent efforts to train machine learning ensembles on war outcomes, which I then use to weight a state's observable material capabilities. I estimate actual international benefits by averaging across a state's centrality in valued international networks. The measure provides multiple advantages over alternative estimates both conceptually and statistically. Beyond its conceptual value, the measure's association with militarized conflict is robust to model specifications, unlike the current go-to measure when modeling country-level sentiments: ideal point estimates from United Nations voting records.

The second chapter asks: when do revisionist states occur? Most responses to this question fall under one of three categories: 1) differential growth rates, 2) domestic political changes, or 3) international dissatisfaction. While these arguments are all based on rich

research traditions, it is an open question as to which theory best predicts when revision occurs. In order to provide such an empirical comparison, I build a series of machine learning ensembles that predict interstate conflict onset and vary only in the included features. While the models for each of the three theories unsurprisingly demonstrate meaningful predictive capacity, the ensemble based on measures of international dissatisfaction is more accurate than counterparts based on differential growth rates and domestic political changes. The paper's results do not invalidate theories of rising powers and revolutionary regimes, but they do emphasize a greater focus on a state's standing within the broader international system.

The final chapter develops a procedure for estimating heterogeneous spillover effects. In social environments, interference between units is likely the norm, not the exception. This poses a problem for estimating causal effects, where the potential outcomes framework assumes that one unit's treatment assignment has no effect on another unit's outcome. In response to this concern, one increasingly popular approach for handling interference between units is the estimation of spillover effects, where sharing a networked tie to a treated unit confers indirect treatment exposure. However, like average treatment effects, there are good reasons to expect that spillovers vary in magnitude and direction across contexts. In order to capture this variation, I approach spillovers through the lens of heterogeneous treatment effects, which can be modeled with causal random forests.

# Acknowledgments

One consistent theme of this dissertation is that relationships matter, in all contexts. My graduate school experience has been no different. Without the support of family, friends, and mentors, this dissertation would never have been possible.

Dan Silverman was the best roommate a first-year graduate student could ever have. He was immensely generous in sharing his hard-earned lessons as he wrapped up his final years of graduate school. Whether it was pushing me to attend conferences, explaining advanced quantitative methods, or walking me through grant applications, I am immensely better off for his friendship and mentorship. Dave Whitsett and Ricardo Graiff-Garcia were the ideal pair of office mates and could easily be considered co-advisors on this dissertation. Miguel Garza-Casado, Kevin Simmt, Liwu Gan, Greg Smith, Ben Campbell, Adam Lauretig, Elias Assaf, Ruthie Pertsis, Jared Edgerton, Soohyun Cho, Aisha Bradshaw, Drew Rosenberg, Austin Knuppe, Reed Kurtz and many other fellow graduate students were instrumental parts of my graduate school experience.

I am immensely thankful for the advisors in my life, both in undergraduate and graduate school. I decided to pursue a Ph.D. after writing my undergraduate thesis, on which Heather Elko-McKibben was a patient and insightful mentor. Randy Siverson gave me invaluable advice about how best to prepare for a Ph.D. program and feedback throughout the application process. While applying to and preparing for graduate school, Jake Hosier was the best boss I could have asked for and endlessly supportive.

At Ohio State I undoubtedly hit the jackpot with my advisors. Bear Braumoeller has been constantly supportive and taught me more than I can express about international politics and statistics. Skyler Cranmer, much like Randy Siverson (who, in the spirit of networks, was a graduate school advisor to Skyler), was clear from the start about navigating the

landscape of graduate school. More importantly, when my father passed away suddenly during my first year of graduate school, Skyler went above and beyond in being present. Chris Gelpi is a model colleague and I find myself almost always asking “How would Chris respond to or interpret this?” I am always better off for doing so. James Wilson has been an invaluable window into the world of statistics and source of positive feedback. Without his insights, this dissertation would be far worse off.

My family and friends have been nothing but supportive of my decision to move across the country to Ohio and pursue a seemingly endless degree. Robert Filipas instilled a sense of self-confidence in me, which was essential throughout the unavoidable roadblocks that come with graduate school. The Bergers are a second family and spending Thanksgivings in Teaneck, New Jersey has been a true highlight these past few years. Bob and Georgia Hamlin’s care packages have spoiled me thoroughly. My childhood group of friends is still going strong and have always been ready to celebrate every step of this journey, for which I am beyond grateful. My mother, Anita, can often be spotted in my hometown proudly sporting her Ohio State gear. Thomas and Ro may be my biggest fans and always lift me up whenever I see them. My father, Cary, is not here to see this, but I think my passion for international politics started with him and playing games like Age of Empires and Stronghold together on our computer growing up. His lessons and insights are surely present.

Lastly, this dissertation is dedicated to Alicia, who also moved to Ohio to be with me throughout this adventure. Her support and partnership is the greatest joy in my life.

## Vita

| | |
|---|---|
| 2013 ..... | B.A. International Relations, University of California, Davis |
| 2017 ..... | M.A Political Science, The Ohio State University |
| 2017-present ..... | Ph.D. Candidate, The Ohio State University |

## Fields of Study

Major Field: Political Science

Studies in:

Quantitative Methods
International Relations

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Measuring International Dissatisfaction

## 1.1 Introduction

A state's level of satisfaction or dissatisfaction with the international status-quo plays a pivotal role in theories of international conflict and cooperation. However, and unfortunately for scholars of international politics, a state's general sentiment toward the international system is directly unobservable, meaning empirical inferences must either be drawn from often-messy observable quantities or avoided altogether. Indeed, insofar as one's focus is a state's dissatisfaction with the international status-quo, the closest available proxy can be found in ideal point estimates derived from United Nations voting records.[1] The popularity of these estimates is unsurprising, given the range of issues that are brought before the United Nations and the related methodological tradition in American Politics around congressional voting records. Moreover, in a dyadic context the estimates are easily interpretable as the ideological distance between two states in a given year, which can then be used as a covariate for any dyadic outcome.

That being said, while ideal point estimates provide a concise comparison of the ideological difference between pairs of states, they are less clearly applicable if one's focus is on any one state's attitude toward the makeup of the international system as a whole, which

[1]The most prominent dataset is found in Bailey et al. (2017). Another set of ideal points are included in Braumoeller (2013), which look at the great powers up to the end of the Cold War and are based on responses to historian surveys.

reflects the aggregation of a state's relations. In response to this methodological challenge, I develop a novel measure of each state's international dissatisfaction from 1816-2012. Drawing on the conceptual framework originally put forward by Gilpin (1983) and more recently applied by Renshon (2016, 2017) to international status, I treat international dissatisfaction as the distance between a state's international *expectations* and the *reality* it faces. The more a state's actual benefits from its international environment fall short of what that state expects it should receive based on power calculations, then the more dissatisfied it will be with the status quo. In terms of statistical modeling, while the exercise is an unsupervised one – we lack consistently matchable outcomes that can statistically validate the final measure – I am able to draw upon popular techniques in network science and machine learning when measuring each relevant component.

Beyond the measure's conceptual distinction, it also provides multiple quantitatively appealing features. First, the United Nations-based ideal point estimates are temporally bounded, spanning only from 1946 to 2012. The dissatisfaction estimates proposed here instead span the entirety of the Correlates of War data set (1816-2012), allowing for a greater range of analytical applications beyond post-WWII and post-Cold War eras. Second, the international dissatisfaction measure is a considerably more robust predictor of whether or not a state initiates international conflict. Fringe ideal points do predict conflict onset, but the magnitude and direction of coefficient estimates is sensitive to model specification with ideal points – which is not the case with international dissatisfaction. Lastly, I find that the relationship between dissatisfaction and conflict initiation is consistent in both statistical significance and effect direction across all years, demonstrating that the relationship is not just driven by a handful of choice outlier cases, as is often a concern around research on international conflict.

Previewing the measure's composition, the final values are produced by estimating and then differencing two component quantities: a state's expected and actual benefits from the international status-quo. Starting with a state's actual benefits, I employ techniques for social network analysis and gather data on a state's position within multiple networks,

where some networks are primarily social and others are primarily material. Conceptually drawing on the relation school in Sociology (e.g., Burt et al., 2005; Emirbayer and Goodwin, 1994; Emirbayer, 1997; Erikson, 2013) and International Relations Theory (e.g., Jackson and Nexon, 1999; MacDonald, 2018; Qin, 2016), I estimate international benefits by calculating each state's position in networks of various valued international goods. These include the: military alliance, interstate trade, shared diplomatic tie, and arms trade networks. Each state's relative centrality in each network is calculated and averaged across all networks in a given year, providing an estimate of each state's yearly access to the systemic distribution of valued international goods – its *actual* international benefits.

Turning to a state's expected benefits, in the Gilpinian framework states become dissatisfied because they believe their share of international goods falls short of what they should be receiving based solely on their power-position. In Gilpin's formulation, expectations are composed of two parts: the possession of and reputation for using material capabilities.[2] The former is relatively straightforward empirically. A state's relative material capabilities can be proxied for by its CINC (composite index of national capability) score, which calculates a state's percent of the globe's total: population, military capacity, resource consumption, and iron and steel production. Measuring the latter Gilpinian estimand – reputation for using these material capabilities – is more difficult and not directly observable. To proxy for a state's material reputation, I replicate and extend Carroll and Kenkel's (2016) recent work, which uses each state in dyad's CINC components to predict militarized interstate dispute (MID) outcomes. After building a machine learning algorithm that accurately predicts which side wins in a MID (ignoring stalemates), I use the model to make predictions about outcomes if all pairs of states in a given year were to engage in a MID. These predicted probabilities are then aggregated and provide general estimates of a state's reputation for using its capabilities. In combination, I then weight a state's observable material capabilities by its estimated reputation for using capabilities, producing a measure that combines material and reputational sources of power.

[2]Gilpin labels this reputation-based component of power 'prestige'.

The final estimate is the difference of these two components – a state's international expectations and reality. The more a state's expected benefits outstrip its actual benefits, then more dissatisfied it will be. Inversely, to the extent that a state's international expectations rest equal to or below their actual benefits, then the more satisfied it will be with the international status quo. The rest of the paper walks through the process of building and validating the measure in the following steps. First, I review the relevant literature. Second, I elaborate upon the my formalization of Gilpin's theory of international dissatisfaction and change. Third, I produce the dissatisfaction measure. Fourth, I test the measure's validity as a predictor of conflict onset and compare its predictive capacity to the aforementioned ideal point estimates.

## 1.2 Literature Review

International dissatisfaction is fundamental to understanding both why change does or doesn't occur and, if change does occur, why it sometimes is peacefully managed versus marred by great power conflict. Whether it be understanding why outcomes vary greatly around rising powers (Allison, 2017; Edelstein, 2017; Fearon, 1995; Gilpin, 1983; Goddard, 2009, 2018a; Goh, 2005, 2013; Kennedy, 2010; MacDonald and Parent, 2018; Organski and Kugler, 1981; Schake, 2017; Shifrinson, 2018; Trachtenberg, 2012), how states seek desired levels of status (Chan, 2004; Duque, 2018; Larson and Shevchenko, 2010; Paul et al., 2014; Renshon, 2016, 2017; Ward, 2017), or debates around the fundamental origins of revisionist states, (Davidson, 2006; Goddard, 2018b; Johnston, 2003; Lyall, 2005; Schweller, 1994, 1999, 2015), actual or potential dissatisfaction with the status-quo is conceptually fundamental. After all, why would an actor change any situation if they are reasonably satisfied with it? Or in the international context, why would a state attempt to alter an environment that appears to be reasonably beneficial? If a state seeks change, then there likely is some aspect of the status-quo they are displeased by and think can be rearranged to be substantially more beneficial, meaning its dissatisfaction is a critical quantity

and merits investigation.

However, despite international dissatisfaction's conceptual importance, it rarely, if ever, receives thorough statistical treatment. This is a particularly strange omission when considered alongside the rich tradition of quantitative analysis in International Relations scholarship. Most likely, the relative lack of such work stems from the simple fact that dissatisfaction is a latent, unmeasurable variable. And, as noted by Lyall (2005), when considering the relevant sample of cases, there is a tendency to only look ex-post at the handful of clear examples of intensely dissatisfied revisionists (e.g., Pre-WWI and WWII Germany, Imperial Japan, Communist Russia, and Revolutionary Iran), which then risks inducing inferential biases through selecting on the dependent variable or ignoring the control group of potential revisionists which did not turn to international conflict and change. In tandem, these methodological challenges suggest that for studies of international dissatisfaction, revision, and change, not only are dissatisfaction or satisfaction unobservable, but gathering an analytically useful sample can be particularly difficult.

Notably, estimates for a state's ideal point are one potential indicator that has grown in recent popularity and can be applied to all cases where yearly data is available. These estimates are most prominently used when based upon voting pattern at the United Nations, where it is assumed that voting records generally represent a state's true preferences (Bailey et al., 2017).[3] Although these ideal points are a valuable contribution to many areas of international politics, they face important methodological and conceptual limitations if one's goal is to study a state's dissatisfaction with the international status-quo as a whole. Methodologically, the data start in 1946, which is a relatively limited time-span historically and does not include some of the most important cases of both revisionist and status-quo states. Conceptually, the estimates place states on points along an abstract ideological spectrum. Quantifying what one location on this ideal point spectrum actually means on its own is a difficult, if not impossible, task. Rather, the estimates primarily provide value if one is interested in a state's general views, *relative to others* based upon the distance between

[3]See Braumoeller (2013) for ideal point estimates for the great powers before 1946 based on historian surveys.

their positions.

While the ideal point estimates allow valuable inferences about how ideologically different pairs or groups of states are from each other, they do not necessarily speak to any single state's attitude toward the international system as a whole. Comparing values certainly sheds light on the similarity or dissimilarity of visions for a best-case international scenario. But comparing values does not necessarily tell us anything about a single state's actual satisfaction or dissatisfaction with their international environment.[4] Put differently, a state's leadership and population may desire some abstract form of the world, but that does not mean anyone actually believes such a world is possible. Only the most powerful states even consider such a world to be in the realm of possibility.[5] Yet, despite the implausibility of actually achieving a best-case scenario, states have consistently taken the side of an international status-quo which they presumably were satisfied by enough to fight for. In this sense, an accurate understanding of any state's attitude toward the status-quo is rooted in something more conservative than ideal visions of some best-case scenario.

One alternative understanding of international preferences (and this paper's focus) can be found in Gilpin's (1983) approach, where dissatisfaction is treated as the difference between a state's international expectations and reality. Gilpin's theory of war and change argues that when states grow more powerful, their desires similarly expand. Should the international environment fail to recalibrate and accommodate these newfound desires, then states tend to turn to war as their preferred mechanism of producing the desired changes. At the core of this conception is the assumption that a state's expected international benefits – labelled as the sum of a its "territorial, political, and economic arrangements."[6] – are a function of relative power.[7] Gilpin breaks power into two categories, the first being

[4] I later find that there is almost zero correlation between a state's ideal point and my dissatisfaction estimates, meaning the two capture different substantive processes.

[5] Though even for the greatest of empires, moments of achieving one's ideal are rarely, if ever, lasting. (Kennedy, 2010)

[6] p11

[7] Gilpin's explicit statement about what constitutes the international status-quo is in itself a valuable contribution. As Johnston (2001) points out, most studies of international dissatisfaction and revision are vague about what revisionists actually seek to change and what it means for them to be dissatisfied. But Gilpin actually lays his cards on the table about what international quantities dissatisfied revisionists are concerned about.

a state's material – primarily military and economic – capabilities. The second, which he labels 'prestige', is a state's reputation for the capacity and willingness to use their material capabilities. Change is sought when a state's power substantially outgrows its share of the aforementioned international goods. Or, as Gilpin argues: "Thus, a precondition for political change lies in a disjuncture between the existing social system and the redistribution of power toward those actors who would benefit most from a change in the system."[8] Generally understood as a occurring through rising powers (though in the next chapter I empirically investigate whether this phenomena applies to all states and empirically compare dissatisfaction to rising powers and domestic changes), Gilpin's precondition for dissatisfaction-driven change is a gap between power and benefits.

The most careful quantitative application of this framework can be found in Renshon (2016, 2017), who applies the expectations-reality approach to a state's desire for international status. When a state's power outstrips its actual status, measured by shared diplomatic ties, then Renshon argues it suffers from a 'status deficit'. Estimates for status deficits are then demonstrated to predict the onset of interstate conflict. Renshon models a state's expected status as its CINC score (Singer et al., 1972) and actual status as its PageRank centrality (Brin and Page, 1998, 2012) in the diplomatic exchange network. This methodological approach is valuable, as it is the first to explicitly take Gilpin's popular theory and demonstrate a way to empirically operationalize the central parts. But it is limited to a single sphere of international benefits (status) and avoids the reputational portion of power. In the next section I take Renshon's general approach and expand upon his methodology to improve in these areas. Put differently, Renshon provides a careful and useful starting point for measuring international dissatisfaction, but only provides a first piece of the broader puzzle.

Lastly, we should note that this exercise is more than just semantics and inside-baseball into the nuts and bots of obscure theories and statistical measures. A Gilpinian understanding of international dissatisfaction is at the core of current analyses of great power politics.

[8] p9

Friedman Lissner and Rapp-Hooper (2018) represent this view well, pointing out that, to the extent the two are antagonistic, Russian and Chinese challenges to the existing international order are rooted in such pressures: "As American hegemony has eroded, so too has the willingness of the United States' near-peer competitors to tolerate a liberal international order which reflects a distribution of benefits that decreasingly resembles the global distribution of economic and military power."[9] In discussions of great power politics, justifications for spoiling and revising the status-quo are often understood or assumed to to be based upon an expectations-reality calculation. However, while conceptually plausible and logical, these analytical claims are generally made based on theoretical expectations. Whether or not they are accurate descriptions of current and past moments, however, is an empirical question and the focus of this paper's remainder.

## 1.3 Model and Theory

Gilpin theorized that the international system is in equilibrium for a member state if the expected costs of international change outweigh the expected benefits.[10] In other words, a state is satisfied with its international environment if they face no incentives to seek significant alternatives. Any potential benefits of change, whether they be territorial, institutional, social, or another, fall short of the expected costs in blood and treasure. This does not mean the status quo is *ideal* for any country, rather no better alternative reasonably exists.[11] Herein, international disequilibrium occurs for any state when its current capabilities, relative to other states, outstrip its relative benefits. Put in less formal terms, a country is unhappy with the current world if their share of the international material pie falls short of what they think they should be receiving according to power-based expectations. If decision-makers find their country in situations where their relative size is greater than their relative share of international goods, then opportunities should exist for to rem-

[9] p13
[10] pp10-11
[11] This is the core difference between United Nations-based ideal point measures and this estimate.

edy this expectations-reality gap by coercive force.[12]

The general equation form of this argument can be understood as:

$$\text{Dissatisfaction}_{it} = \text{Capabilities}_{it} - \text{Benefits}_{it} \tag{1.1}$$

where $i$ is an individual state and $t$ a given year. This formulation brings forward at least one immediate question: What does it mean for a country to receive *international benefits*?[13] A state's well-being is linked to prosperity and security, which of course are at least partly product of domestic resources such as institutional design, leadership, and more. But the international system also provides opportunities for improving a state's strategic position and general prosperity. Countries benefit from trade, diplomatic exchanges, alliances, and more. In other words, the relations that constitute international relations are a source of international benefit. While functioning domestic institutions and markets are necessary for a country to exert itself internationally, so too are robust relations with other states. International politics are not just conducted under the shadow of security concerns; states are also concerned with making the most out of opportunities to prosper through relations with each other. [14]

I operationalize a state's international benefits as the aggregation of its strategic relations with other states. However, rather than simply summing the number of a state's relations across a network in a given year, I weigh relations with more important states more heavily. I do so by measuring each state's PageRank centrality in its international network. (Brin and Page, 1998, 2012) PageRank centrality aggregates the number of ties to a node in a network, but it weighs ties to more central nodes more heavily than ties to nodes

[12]Admittedly, this runs into Fearon's (1995) puzzle of why countries would ever need to turn to war, when war is costly and a preferable war-avoiding negotiated settlement should exist. However, in this context if a disequilibrium persists, then I assume that for some reason the relevant states have been unable to reach a war-avoiding negotiation. If this were possible, the incentives to avoid costly war should have kicked in and avoided the disequilibrium altogether. That being said, the mechanics of why states might fail to reach these negotiations in specific cases are not the focus for this paper. Why states might fail to accommodate these *systemic* imbalances is an open and interesting question for future work.

[13]The determinants of a country's international capabilities are certainly far from obvious, but are a less abstract concept than international benefits.

[14]This class of argument draws heavily upon the relational-network approach to social systems. For more on this type of theory see Burt et al. (2005); Erikson (2013); Emirbayer and Goodwin (1994); Jackson and Nexon (1999); Hafner-Burton et al. (2009); Wasserman and Faust (1994)

that are less central.[15] The underlying principle in its estimation process is that a node is more central the more connections it has, and connections are more important when they are made to other highly connected nodes. Formally, PageRank centrality is expressed as:

$$x_i = \alpha_A \sum_j a_{ij} \frac{x_j}{g_j} + (1 - \alpha_A) \frac{1}{N} \tag{1.2}$$

where $g_j = \sum_i a_{ji}$ (node $j$'s total number of ties). $x_i$ is the PageRank centrality for node $i$ and $x_j$ is the centrality for node $j$.[16] $a_{ij}$ is equal to 1 if a tie exists between nodes $i$ and $j$, but it is equal to 0 otherwise. $\alpha_A$ is a constant damping factor that weighs how much a node's ties matter, as opposed to treating each node as equally central (if $\alpha_A = 1$); in the context of Google's search engine the damping factor approximates the probability that a user will stop clicking links at any time. A visual example of PageRank centrality applied to an example from international politics can be found in Figure 1.1.

Once estimated, we can take each state's centrality estimate across networks and estimate total benefits by averaging:

$$\text{Actual Benefits}_{it} = \frac{1}{N} \sum_{n=1}^{N} \text{PageRank}_{int} \tag{1.3}$$

with $\text{PageRank}_{int}$ referring to state $i$'s PageRank centrality in year $t$ in a network $n$, which is then averaged to provide a state's general access to valued international goods in a given year. The aggregated benefit estimates across all states in a given year end up summing to 1, meaning each state's benefits are a percentage, which is useful for inferential purposes. I subsequently outline the data used in more detail, but the networks included for this measure are: alliances, trade, shared international organizational membership, shared diplomats, arms-trading, and defense cooperation agreements.

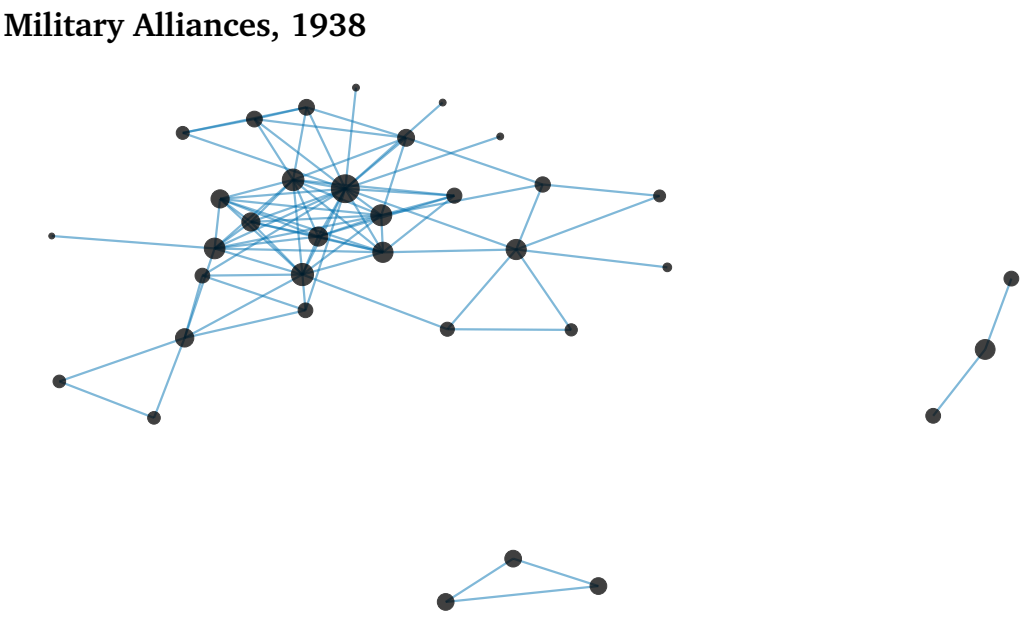
Turning to a state's capabilities, Gilpin theorizes that a state's power in any given year is a function of *observed capabilities* and *reputation for resolve and capabilities*.[17] Gilpin labels

[15] The original idea was that the internet can be viewed as a network where nodes are sites and ties are links to one site that are included on another site. Google's search engine then ranks results based upon a version of PageRank centrality, where a site's ranking on the search algorithm is based upon its PageRank centrality.

[16] If one carefully reads the formula, then they will see that calculating a node's PageRank centrality requires having already estimated every other node's PageRank centrality. This produces a chicken-egg problem where there is no obvious answer about how to estimate initial values. For a mathematically detailed explanation of how this problem is overcome see: http://www.ams.org/publicoutreach/feature-column/fcarc-pagerank.

[17] Gilpin labels these as capabilities and prestige.

Figure 1.1: PageRank Example: Interstate Military Alliance Network, 1938




The interstate military alliance network in 1938, before the outbreak of World War II. Node size is a state's PageRank centrality, with larger nodes representing more central nodes. Each tie represents the presence of a military alliance between two states (either a defensive alliance, offensive alliance, or neutrality pact). Country labels are not intentionally left out to not distract from varying node sizes. Isolates (nodes with no ties) are intentionally left out as well.

the latter "prestige", elaborating: "Prestige is the reputation for power, and military power in particular. Whereas power refers to the economic, military, and related capabilities of a state, prestige refers primarily to the perceptions of other states with respect to a state's capacities and its ability and willingness to exercise its power."[18] In other words, a state's power is a function of both its aggregated oversable material capabilities *and* other states' perception of a state's ability and willingness to use those capabilities. Moreover, while the former – a state's observable capabilities – has received a good deal of attention in the literature and reasonable measures are readily available, the latter – a state's reputation for its material capabilities – requires more careful treatment in the subsequent pages, where

[18] p.31

it is inferred from MID outcomes.[19]

I separately produce estimates for observed capabilities and perceptions of state capabilities and then combine the two into a single measure:

$$\text{Capabilities}_{it} = \text{CINC}_{it} \times \frac{\sum_{j=1}^{J-1} \Pr(\text{Win}_j)}{J-1} \tag{1.4}$$

Where $\text{CINC}_{it}$ refers to the composite index of national capabilities, a measure of a state's percent of all material goods for a given year. CINC scores capture the observed portion of Gilpin's conceptualization. The subsequent perceptions-based portion of the equation $\frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$ is more complicated. Extending Carroll and Kenkel (2016), I build a machine learning ensemble to predict the outcome of militarized interstate disputes (MIDs) that have taken place between countries. Then I take the predictive ensemble and use it to predict the probability of each state winning a MID *if every dyad from 1816-2012 were to engage in a MID*. These predicted probabilities can then be aggregated to provide us with a relative sense of each state's expectations about its capabilities and those of others – a proxy for Gilpin's perception-based quantity. Turning to the equation, for every state in every year, I average each states' probability of winning across all possible MIDs in a given year. The aggregated probabilities are expressed through $\sum_{j=1}^{J} \Pr(\text{Win}_j)$, where $J$ is the number of states in a given year (I subtract 1 because a country does not fight itself in interstate conflict) and $\text{Win}_j = 1$ if state $i$ expects a victory against state $j$; summing to the number of expected victories. I then divide by $J-1$ to standardize between 0 and 1, like the international benefits estimates.

Before combining the two components into a final dissatisfaction measure, I implement a smoothing function on both to minimize the presence of inaccurate spikes due to measurement issues. For each state, the smoothing function takes all input values from a previous number of specified years and computes an average. The formulation is as follows:

[19] One option is to understand reputation for use of material capabilities as a state's *resolve*. For a micro-oriented treatment of resolve in international politics, see Kertzer (2016).

$$\mathcal{S}(x_{it}) = \frac{\sum_{t-n}^{t}(x_{it})}{n} \tag{1.5}$$

where $x_{it}$ is the input value for state $i$ at time $t$, $n$ is the number of time periods considered for the smoothing function[20], and $\mathcal{S}$ represents the smoothing function itself. Past years only are considered for the smoothing function because conflict onset – which I theorize is related to dissatisfaction – can influence a country's subsequent dissatisfaction, either furthering social exclusion or successfully remedying dissatisfaction-inducing grievances.

In sum, combining the two components gives us our final measure:

$$\text{Dissatisfaction}_{it} = \log\left[\mathcal{S}\left(\text{CINC}_{it} \times \frac{\sum_{j=1}^{J}\Pr(\text{Win}_j)}{J-1}\right)\right] - \log\left[\mathcal{S}\left(\frac{1}{N}\sum_{i=1}^{N}\text{PageRank}_{int}\right)\right] \tag{1.6}$$

where $\text{Dissatisfaction}_{it}$ is state $i$'s international dissatisfaction with the status quo in a given year $t$. $\mathcal{S}\left(\text{CINC}_{it} \times \frac{\sum_{j=1}^{J}\Pr(\text{Win}_j)}{J-1}\right)$ is a state's smoothed expected international benefits in year $t$ and $\mathcal{S}\left(\frac{1}{N}\sum_{i=1}^{N}\text{PageRank}_{int}\right)$ is a state's smoothed actual international benefits in year $t$. I log both components because they are heavily right skewed, with most states being low in capabilities and low in centrality. Much like the common practice in Economics of logging right-skewed variables, such as an individual's income, to produce a relatively normal distribution, this helps us parse out more fine-grained variation across states for each component. On the final scale positive values correspond to dissatisfaction and negative values to satisfaction. The theory then predicts that the more dissatisfied a state is (increasingly positive values), then the more likely that state is to start an interstate conflict. Dissatisfied states believe they are being materially and socially shorted by the international system. These states then are expected to turn to militarized coercion in order to remedy their dissatisfaction.

[20] I settle on 4 years for each smoothing function.

## 1.4 Data

### 1.4.1 Expected Benefits

Each measure – a state's relative capabilities and its access to international goods – is a composite of multiple data sources. Breaking down each portion of the relative capabilities measure, recall that a state's relative capabilities are estimated as:

$$\text{CINC}_{it} \times \frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$$

and serve as a combination of observable capabilities and a state's reputation for using those capabilities. The observed portion of capabilities are a state's composite index of national capabilities (CINC) score (cites) for a given year. Started by Singer et al. (1972), a CINC score represents a state's percent of the globe's material capabilities in a given year. The variables used relate to: industrial capacity, population, wealth, and military size.

While the CINC score captures readily observable quantities, the equation's second portion: $\frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$ is meant to approximate a state's reputation for its use of capabilities. This is the perceptions-based portion of Gilpin's formulation, where, building on work by Carroll and Kenkel (2016), I train a machine learning ensemble on the components of each state's CINC score to predict the outcome of past militarized interstate disputes. Once a reliable model is trained on the conflicts that have occurred – where the model is evaluated using a test set of data not included in training – the model is used to make predictions for the outcome of all possible MIDs from 1816-2012. This provides predictions about how many MIDs each state could win, given all possible MIDs, in a given year. Although certainly not the exact same as a state's reputation for using force, a model which predicts conflict outcomes well (both in and out-of sample) serves as a reliable proxy and best guess for what would occur if any two pairs of states were to fight. Importantly, the two quantities appear to capture separate phenomena, with a correlation of 0.469 between the two.

Notably, I have to address one substantial issue when adapting the DOE scores to my

purposes. Most MIDs end in a stalemeate. Indeed, in the DOE dataset 84% of MIDs end in a stalemate. This likely occurs because states only select into conflicts if they think they have a chance of winning. However, in practice, this means that when predictions are made about hypothetical MIDs, such as the United States vs. Fiji, the prediction is that the hypothetical MID will end in a stalemate. Although hypothetically a selection process could occur where a MID would only break out between the two if Fiji believed it had a chance of standing up against the U.S. military, a prediction which treats the two as equally capable is an extreme stretch. Ultimately the original DOE estimates predict that essentially all possible dyadic conflicts from the end of World War II-forward will end in stalemates. If the goal is estimate *relative* capabilities, then a model that considers all states to be *equally* powerful is not necessarily useful.

To avoid this issue, I shrink the training set substantially and only consider MIDs where the outcome is not a stalemate. The final predictions then consider: *if each dyad were to fight and one side were to win, then how likely is each side to be the victor?* Although this lowers us to a training set of n = 270, the machine learning algorithm is able to converge upon a reasonable fit both in training and test set. Indeed, in the the test set, where prediction is inherently more difficult, 72.3% of cases are predicted correctly. Lastly, turning back to the formula of interest – $\frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$ – we can sum a state's predicted probability of winning a MID for each possible dyad in a given year and then divide by the number of pairs to get an average probability of victory. I summarize the predictions in the next section.

### 1.4.2 Actual Benefits

The actual benefits portion of international dissatisfaction is estimated as:

$$\text{Benefits}_{it} = \frac{1}{N} \sum_{i=1}^{N} \text{PageRank}_{int}$$

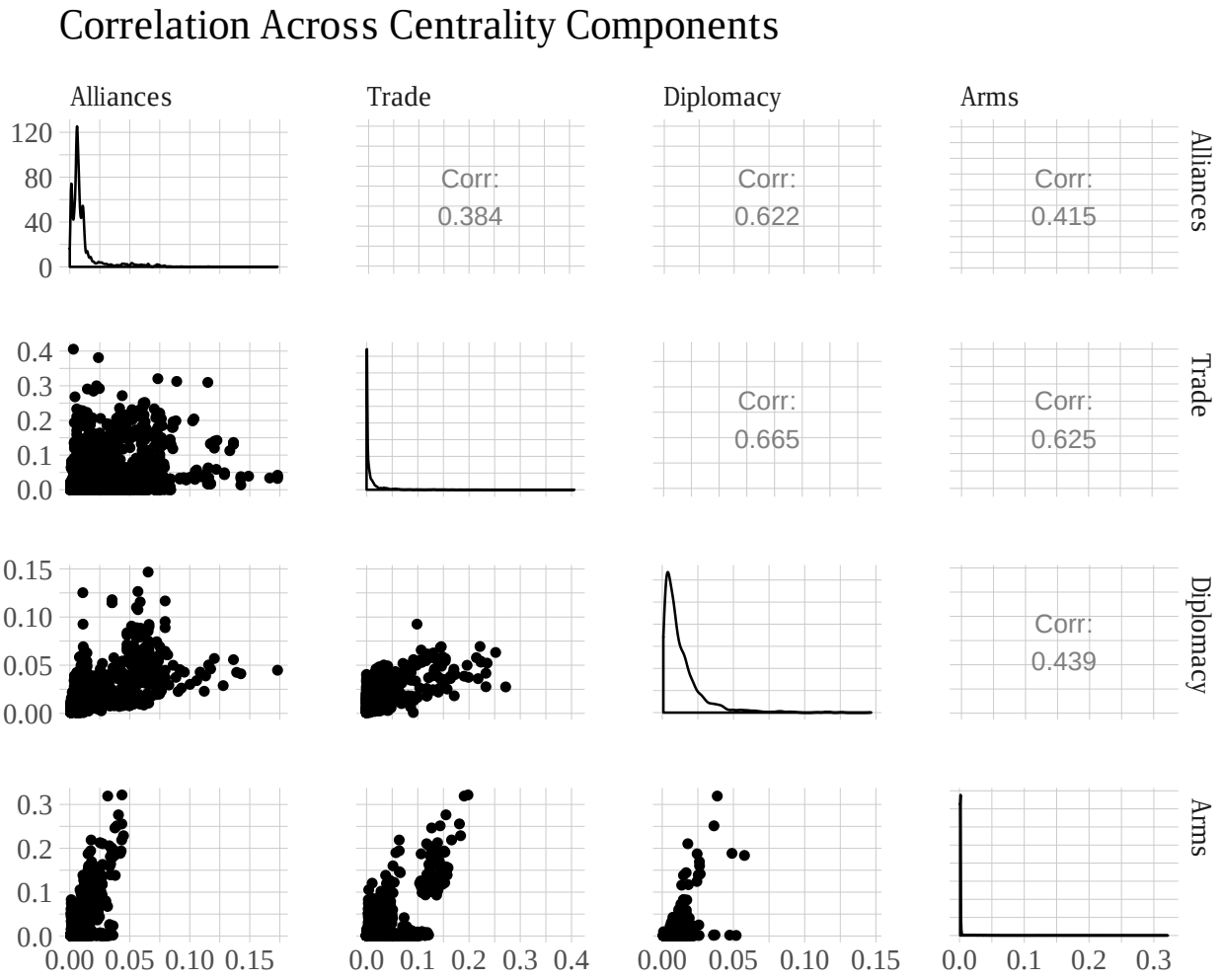
Where $\text{PageRank}_{int}$ refers to a state's PageRank centrality in a network in a given year, which is then aggregated across every available network and averaged. This gives us a

measure of how central a country is across networks, on average. Here I opt for parsimony and only included variables which are available across most, if not all, states over time and space. These are a state's: military alliances, interstate trade, shared diplomatic relations, and arms transfers. The most notable exclusion from this list is shared membership in international organizations. However, and unfortunately for inference purposes, most states are members of most organizations. So the network of shared organizational memberships is too dense to reliably parse out which states are more central than others. While the trade network is also dense, the weight of ties (the amount of trade between two countries) differs enough across dyads to consistently extract meaningful variation.

Estimates for a state's centrality within the international alliance network are produced using the Alliance Treaty Obligations and Provisions (ATOP) dataset (Leeds et al., 2002). Trade data comes from the correlates of war interstate trade dataset (Barbieri et al., 2009; Barbieri and Keshk, 2016), shared diplomatic ties are a combination of data from the correlates of war diplomatic exchange dataset (Bayer, 2006) and data used in Duque (2018), and arms transfers were downloaded from the SIPRI conventional arms transfers dataset (Stockholm International Peace Research Institute, 2019). The datasets vary in their temporal span, with the alliance dataset being the most expansive, covering the entirety of the estimates (1815-2012). See Figure 1.2 for a correlation matrix with each component centrality variable.

Returning to the earlier discussion of Renshon's (2016; 2017) arguments about status dissatisfaction, while the diplomacy network (which is the sole focus of Renshon's analyses) displays a correlation with the other networks (0.622 and 0.665, respectively), there is a substantial swath of variation which it does not capture. Indeed, a closer look at the bottom-left portion of the matrix suggests that much of the correlation across all variables stems from most countries being low in centrality within all networks. Like many social processes the diagonal elements – which contain density plots of the distribution for each variable – are heavily right skewed, with fat-tails. (Barabási and Albert, 1999; Clauset et al., 2009) But as countries reach farther from zero in centrality for each network, then their centrality

Figure 1.2: Correlation Matrix of Centrality Components




Correlation matrix of each component included in a state's average centrality score. In the bottom-left plots, each point is a state's centrality measure for the variable of interest in a given year. Diagonal elements are a density plot of the distribution of each variable for all country-year combinations. The top-right plots list the correlation between each variable combination. While there is some correlation across variables – it would strange if there were not – the highest correlation between variables is at 0.665 between diplomatic ties and trade. Each variable likely is picking up substantial independent variation.

in other networks starts to diverge. In this sense, particularly among the great powers that tend to have the capability to carry out genuine systemic revision, one's position within the diplomacy network is only a partial indicator of one's position within the trade, alliance, and arms-transfer networks.

## 1.5 Results

In this section I, in the following order, introduce component estimates, the final estimates, and compare statistical models of interstates conflict onset based upon ideal point estimates and the dissatisfaction measure. Models are fit to estimate both: whether a state initiated any militarized interstate disputes (MIDs) and the number of initiated MIDs by a state in a given year. Across model specifications, the dissatisfaction measure maintains a positive and significant relationship with a state's propensity to initiate MIDs Extreme ideal points are also associated conflict onset, but the relationship is far less consistent across models – flipping directions and varying in statistical significance depending upon model specification.

### 1.5.1 Expected Benefits

Starting with the expectations-based component of a state's international dissatisfaction, the main challenge is producing the expectations-based component of material capabilities. A countries observable material capabilities are simply treated as its CINC score in a given year, but the expectations-based component are an extension of Carroll and Kenkel's (2016) 'Dispute Outcome Expectations' scores, which are the predicted probability in all possible dyads that State A wins, State B wins, or as stalemate occurs. Unfortunately, due to patterns in MID outcomes, almost all dyads from around World War II-forward are predicted as almost-guaranteed stalemates. This limits the utility of the original DOE estimates, so I update them to only be trained on MIDs where there is no stalemate. While this limits the size of the training set, which always risks removing informative variation, in this case it ultimately provides more useful estimates. These estimates can therefore be treated as the predicted probability that one side will get a better deal in a war-ending bargain than the other side.

I produce these estimates through an automated machine learning (AutoML) algorithm,

a process developed by H2O artificial intelligence research team.[21] In the AutoML algorithm, the following models are fit to a specified training set across potential hyperparameter combinations: a random forest, an extremely-randomized random forest, a grid of gradient boosting machines, a grid of deep neural nets, a grid of GLMs, and two stacked ensembles.[22] (The H2O.ai team, 2015) Models are fit with k-fold cross-validation and then their training performance (here with PR-AUC and AUC-ROC since the outcome is a binary catgory) is stored so the user can decide which of the available models to use for predictions, with the best-performing model (or leader) being the default choice.

H2O's algorithm is admittedly not the only AutoML option (other popular options include, but are not limited to, AutoSklearn, AutoKeras, and Darwin). This raises the question of why I choose H2O's software over other options. Ultimately, while no single AutoML software is universally agreed to outperform all others, recent research suggests that H2O's algorithm is equally as effective as, if not marginally more effective than, any other AutoML algorithms at binary classification tasks such as this exercise. (Truong et al., 2019) In this case, after fitting various models to the training set, the AutoML algorithm settles upon a gradient boosting machine.[23] (Friedman, 2001, 2002) It is not necessarily surprising that the AutoML algorithm settles upon a tree-based model, rather than a deep neural network (which are often considered the state-of-the-art for predictive modeling), because neural networks are generally understood to require much more data than is available in a dataset like the population of MIDs. Turning to the test set, the algorithm predicts approximately 72% of all out-of-sample MID outcomes correctly. Test set predictions and actual outcomes are compared below in Table 1.1.

For a more fine-grained breakdown of the CINC scores, machine learning estimates, and composite measure over time, see Table A.4 in the appendix.[24] In Figure 1.3, I present the

[21] https://www.h2o.ai/

[22] A simple overview can be found http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

[23] Technically, the AutoML algorithm settles upon an XGBoost model, which is a type of gradient boosting machines. (Chen and Guestrin, 2016)

[24] Tables include the average CINC scores, average probability of wining a hypothetical MID, and the average estimated expected international benefits for the top-10 highest expecting states in the following time periods: Pre-WWI, Interwar, Cold War, and Post-Cold War.

Table 1.1: Confusion Matrix of Predicted and Actual Outcomes

| | | Actual | | |
|---|---|---|---|---|
| | | Victory A | Victory B | Accuracy |
| Prediction | Victory A | 20 | 7 | 0.74 |
| | Victory B | 4 | 10 | 0.71 |
| | Total | 24 | 17 | 0.73 |

Confusion matrix of predicted and actual outcomes in the *test set*. Training set performance is nearly perfect, though is this is not an indicator of model accuracy because of the tendency of machine learning models to overfit. (Neunhoeffer and Sternberg, 2019) Because the model is trained on dyadic conflicts where one side is the clear winner, Victor A and B represent which of two states is predicted to be or actually is the victor. Elements in the same row and column ((Victory A, Victory A) and (Victory B, Victory B)) represent accurate test set predictions.

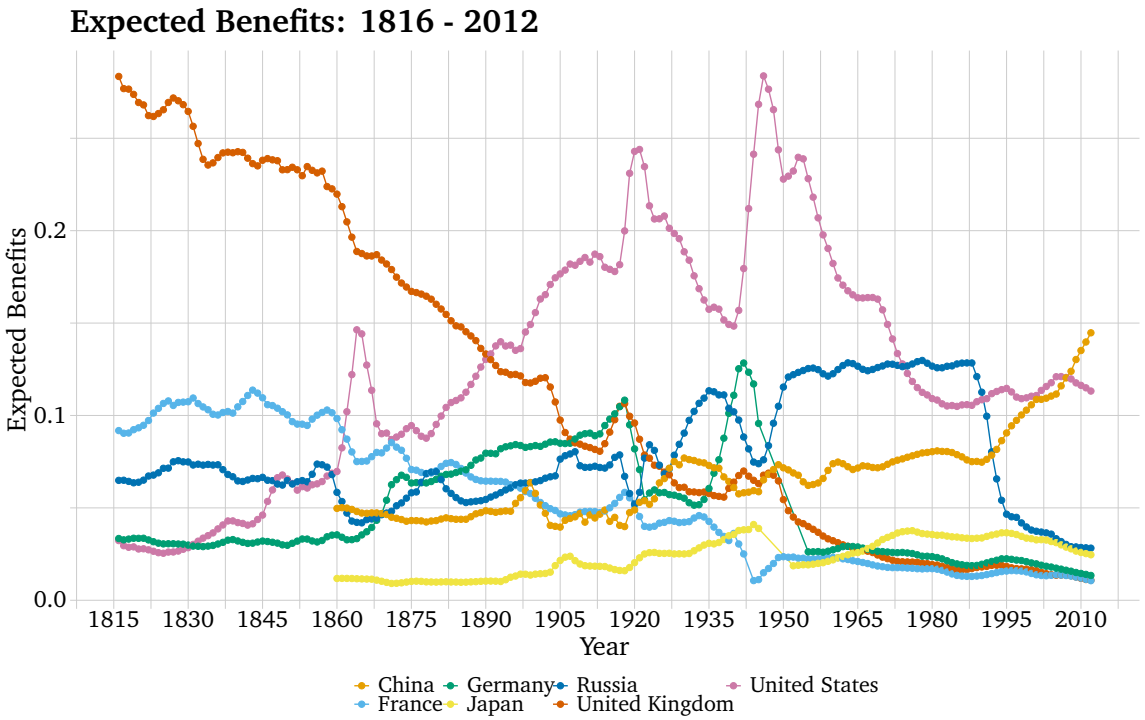
estimated expected international benefits for various great powers over time. The estimates generally follow conventional understandings of power rankings in the literature, but with some important nuance. The United Kingdom initially starts with the greatest expectations, with a shift around the United States power transition in the 1890s. Germany spikes in expectations before both World Wars. During the Cold War there is a moment of Russian surpassing the United States.[25] The post-Cold War era is initially characterized by United States dominance, but we see China overtaking the US in its desires near the end of the time series. Russia's modern expectations fall considerably short of both the United States and China.

### 1.5.2 Actual Benefits

Second, let's consider the measure's actual benefits component. To reiterate, I estimate a state's access to valued goods within international networks and then take the average of each centrality score. While the expected benefits component was estimated based upon a iterative process of tuning a machine learning algorithm until test set predictions were sufficiently accurate, the centrality estimates were more straightforward to estimates because

[25]Indeed, this captures the general sense that during the Cold War it was far from obvious which side was more powerful.

Figure 1.3: Expected Benefits Estimates



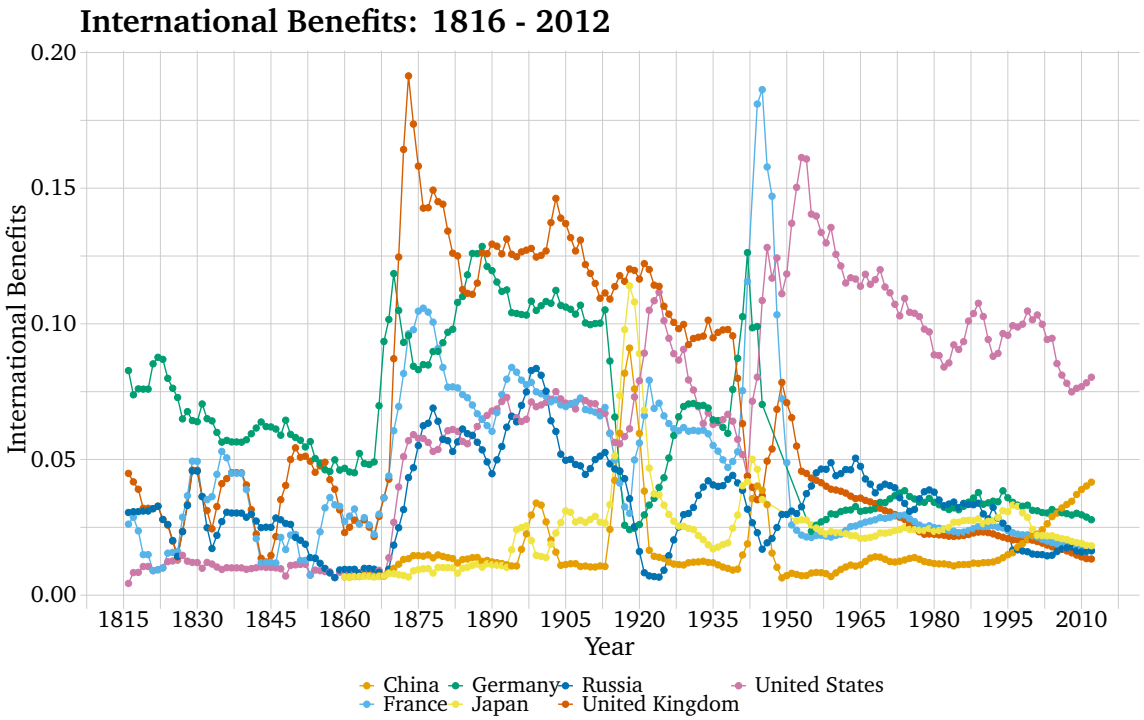

Estimates for a state's *expected* international benefits from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable.

I use an already-established centrality formula.

The estimates are included in Figure 1.4. Like Figure 1.3, each point represents a country's actual estimate for a given year. The y-axis includes decimal values because estimates represent the percent of global benefits that any single state accesses in a given year. Much of the data follows general understandings of international history, with the United Kingdom and United States receiving the majority of international benefits for much of the data. Interestingly, we see China rising in terms of benefits from the early-1990's forward. But China's portion of international benefits in 2012 falls substantially short of the United States. While China is a greater beneficiary than the other included great powers, the gap between China and the United States illustrates an important point. During the last available estimate (2012), The United States nearly doubles China's estimates.

On this note, the gap between the United States and China in estimated international

Figure 1.4: Actual Benefits Estimates



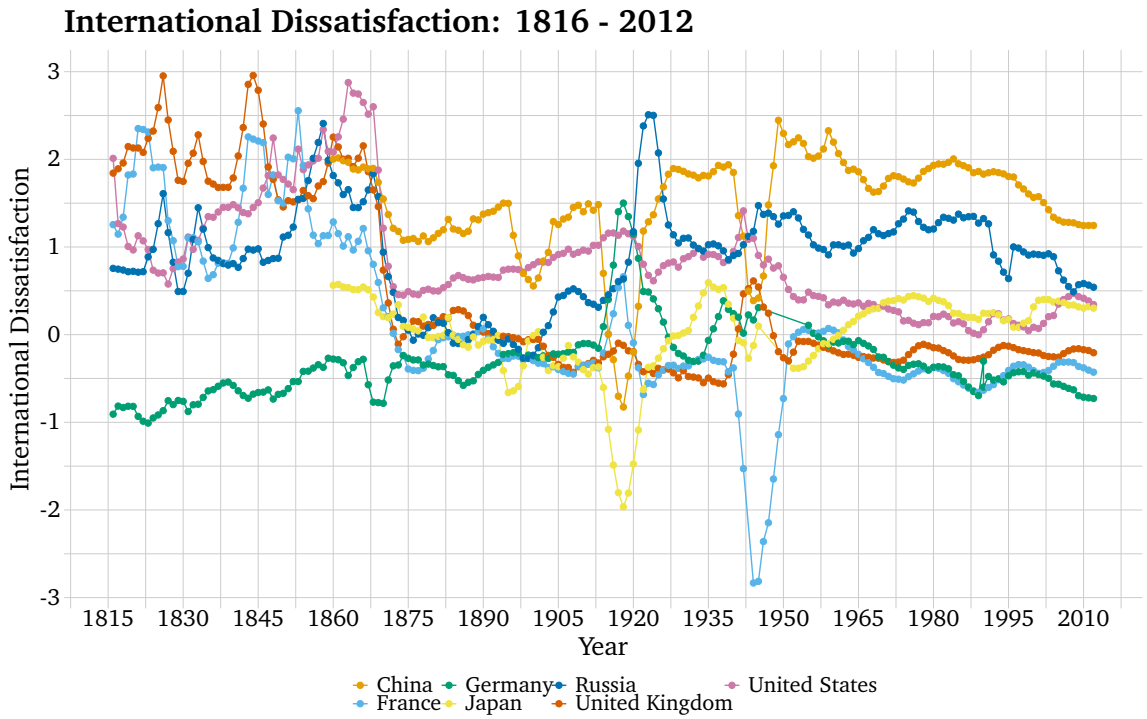

Estimates for a state's *actual* international benefits from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable.

benefits sheds light on one important question regarding China's rise. When it comes to the debate about whether China's growth poses a revisionist threat, a popular question is to ask why a country would alter a system which benefits it? While China has reason to consider itself as powerful, if not more powerful than the United States, it lags behind in actual benefits from the international system. In this sense, while the international system may be growing increasingly beneficial for China, that does not mean the international system is sufficiently beneficial, nor is guaranteed to be so, in China's eyes. Turning to the final estimates, we see that this has been the case for some time.

### 1.5.3 Dissatisfaction

Figure 1.5 includes the final dissatisfaction estimates for select great powers. The figure's format is the same as the preceding figures, where the y-axis includes the measure, the x-axis is the year, and each state is represented by a different color. Positive values on the y-axis correspond to increasingly *dissatisfied* countries. Negative values on the y-axis correspond to increasingly *satisfied* countries.

Figure 1.5: International Dissatisfaction Estimates




Estimates for a state's international dissatisfaction from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable. Colors correspond to the country. Positive values on the y-axis correspond to increasingly dissatisfied states. Negative values on the y-axis correspond to increasingly *satisfied* states. Note that before each World War Germany spikes in dissatisfaction.

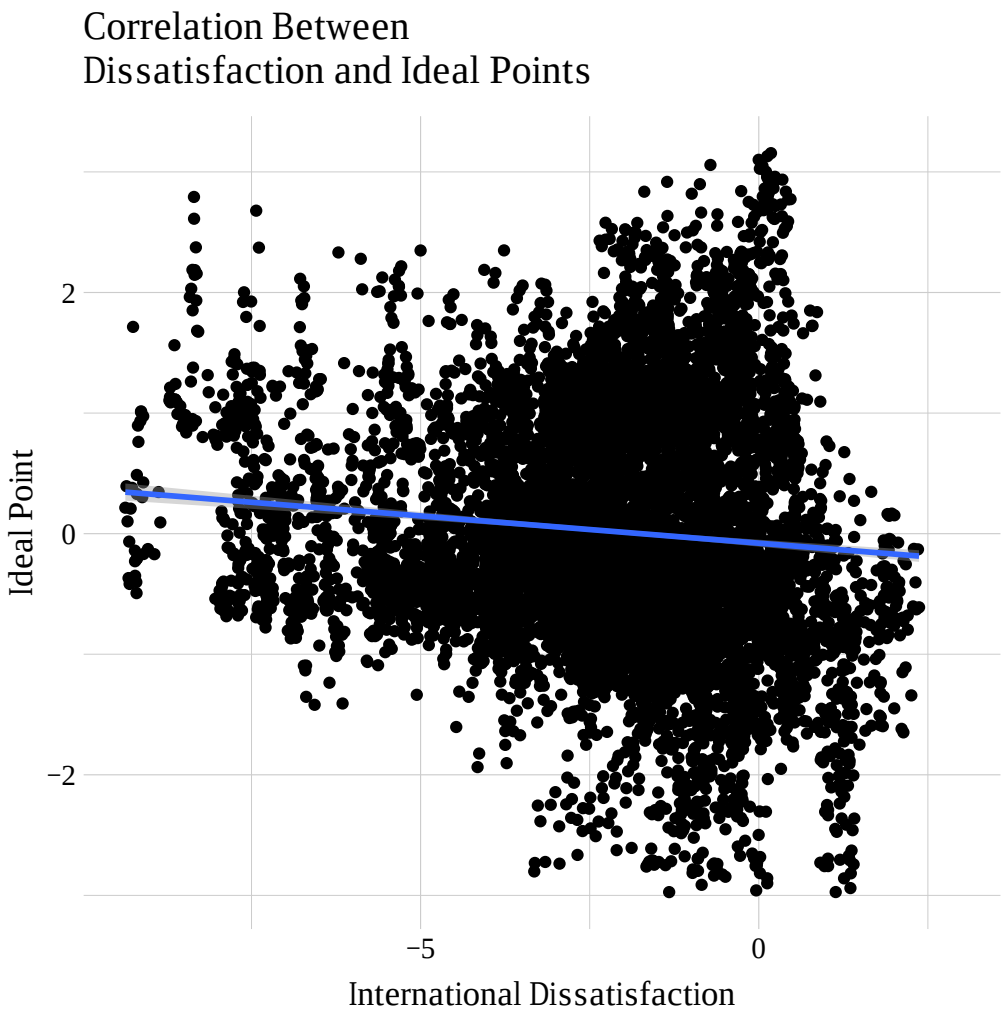
Considering the two World Wars, which are given understandable focus in study of international dissatisfaction and revision, Germany spikes in dissatisfaction right before

both the start of both wars. Turning to the modern era, China and Russia are the two most dissatisfied of the great powers. Russia's dissatisfaction is estimated as being in decline, but we can understand this in the context of Russia's general economic decline. While still a formidable world force, its post-Cold War status is not what it once was. Therefore, we should expect its international expectations to be in decline. Moreover, the European great powers – France, Germany, and the United Kingdom – are all satisfied, which is unsurprising but evidence in support of the measure.

Two surprising outcomes are Japan and the United States both being somewhat dissatisfied in the modern era. Though surprising, these are also the types of estimates that make the modeling exercise worth doing. If the estimates perfectly meshed with general historical narratives, then there would be little added value. At first glance, these results are not necessarily the narrative one would immediately prescribe to both countries; the United States sits atop the international hierarchy and Japan's economic and cultural standing are well-documented. However, the United States is arguably the most powerful country in world history. Though that has translated to international hegemony, it does not mean its post-Cold War benefits are commensurate to its power-based expectations, which are estimated to fall short. Similarly, Japan is a world economic power, but a stagnant level of international benefits may be a legacy of its role in World War II and earlier wars, making peer states wary of Japan's potential for revisionist tendencies.

Lastly, this measure was presented as substantially different from a ideal point measures. Is this borne out? If, after all of these estimation routines, we just end up with similar values reached through different methods and data, then the exercise has little value. Fortunately, that is not the case. Figure 1.6 plots ideal point estimates against the final dissatisfaction values. The correlation between the two is almost zero (actually at -0.094), meaning there is almost no association between the two variables when they are plotted against each other. The lack of association between the two variables gives us confidence that the modeling exercise is substantively useful and captures a distinct quantity.

Figure 1.6: Correlation Between Ideal Points and Dissatisfaction




Correlation between United Nations-based ideal point estimates and international dissatisfaction. Each point is a country-year. The correlation between the two is -0.0946, which is demonstrated by the regression line along the plot.

### 1.5.4 Statistical Performance

How can we deem the dissatisfaction measure a valid statistical construct? According to the Gilpinian theory of international dissatisfaction, the more dissatisfied a country is, then the more we should expect it to turn to coercive military force as a means of remedying that

gap. The theory itself is agnostic about which country(ies) will be targeted. But some form of militarized force is expected to follow. Moreover, beyond whether or not the measure has some predictive capacity, we also care about how well it does relative to a popular alternative. At the outset of the paper, I situated international dissatisfaction in comparison to a state's ideal point estimate, based on its United Nations voting records. In the following statistical models I estimate a logistic regression of whether or not a country initiated a militarized interstate dispute in a given year and then a poisson regression of the number of MIDS initiated by a country in a given year.

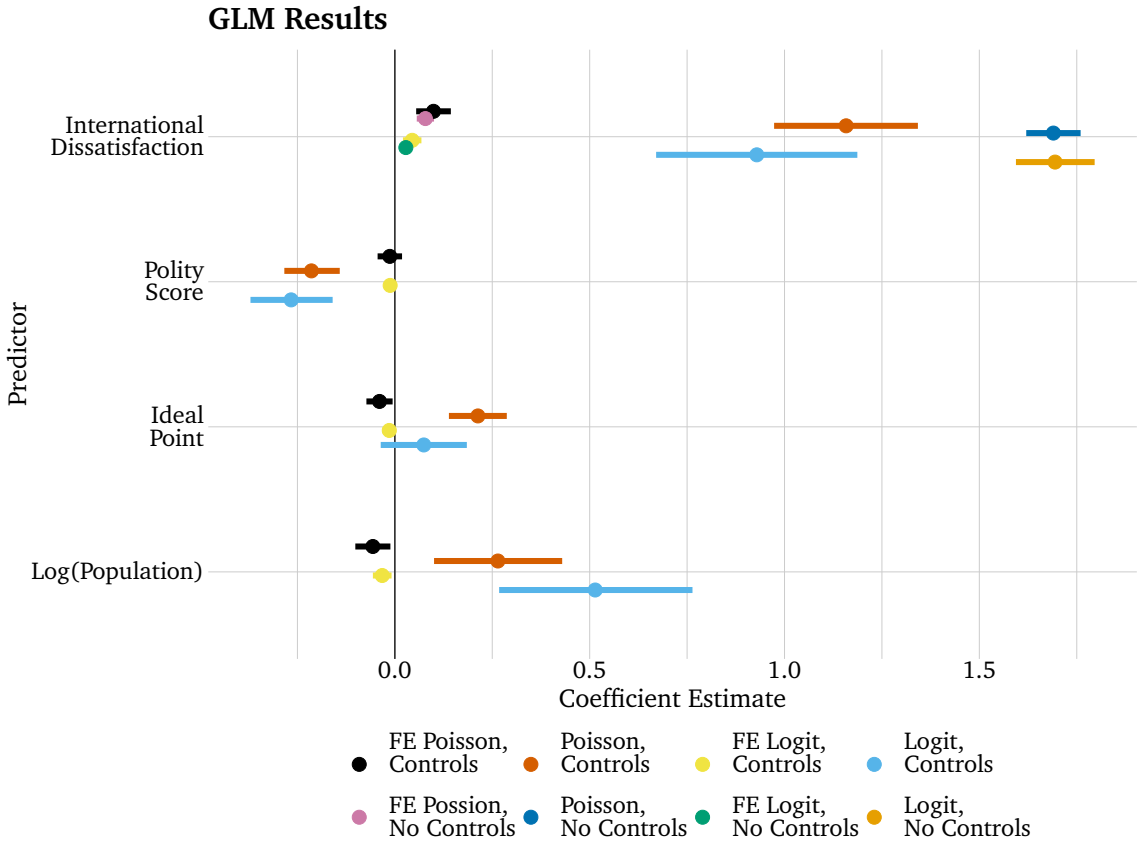
The general formulation is as follows:

$$\text{MID}_{it} = \alpha_i + \gamma_t + \beta \cdot \text{Dissatisfaction}_{it} + \eta \cdot \text{Ideal Point}_{it} + \delta \cdot \text{Control Variables}_{it} + \epsilon_{it} \quad (1.7)$$

where $\text{MID}_{it}$ is whether (or the count) a state $i$ initiates a MID in year $t$, $\alpha_i$ are individual-level fixed effects, and $\gamma_t$ are year effects. For control variables, while a standard set of controls are considered in dyadic studies of conflict (joint democracy, peace years, distance, etc.), a standard set of controls for a monadic study such as this one are less apparent. Following Braumoeller (2019), I only consider reciprocated MIDs as the dependent variable. Combined with the new MID coding by Gibler et al. (2016), this gives us a population of MIDs that does not involve completely inconsequential MIDs but includes the low-level MIDs that could very well have escalated further than they did (reciprocated). I also control for a state's polity score and its population, as one can plausibly see a situation in which more autocratic states are both more dissatisfied and have a greater tendency for initiating conflict. Population is also potentially prior to both conflict and dissatisfaction. I therefore use the logarithm of a country's population as a proxy for size instead of CINC scores, because of the prominent role that CINC scores play in the international dissatisfaction measure. Lastly, fixed effects are included in some models in order to soak up other time-invariant unobservable confounders. Models with any conflict onset as the dependent variable are a logistic regression and models with the number of expected conflicts initiated

are a poisson regression.

Figure 1.7: GLM Coefficient Estimates and Standard Errors



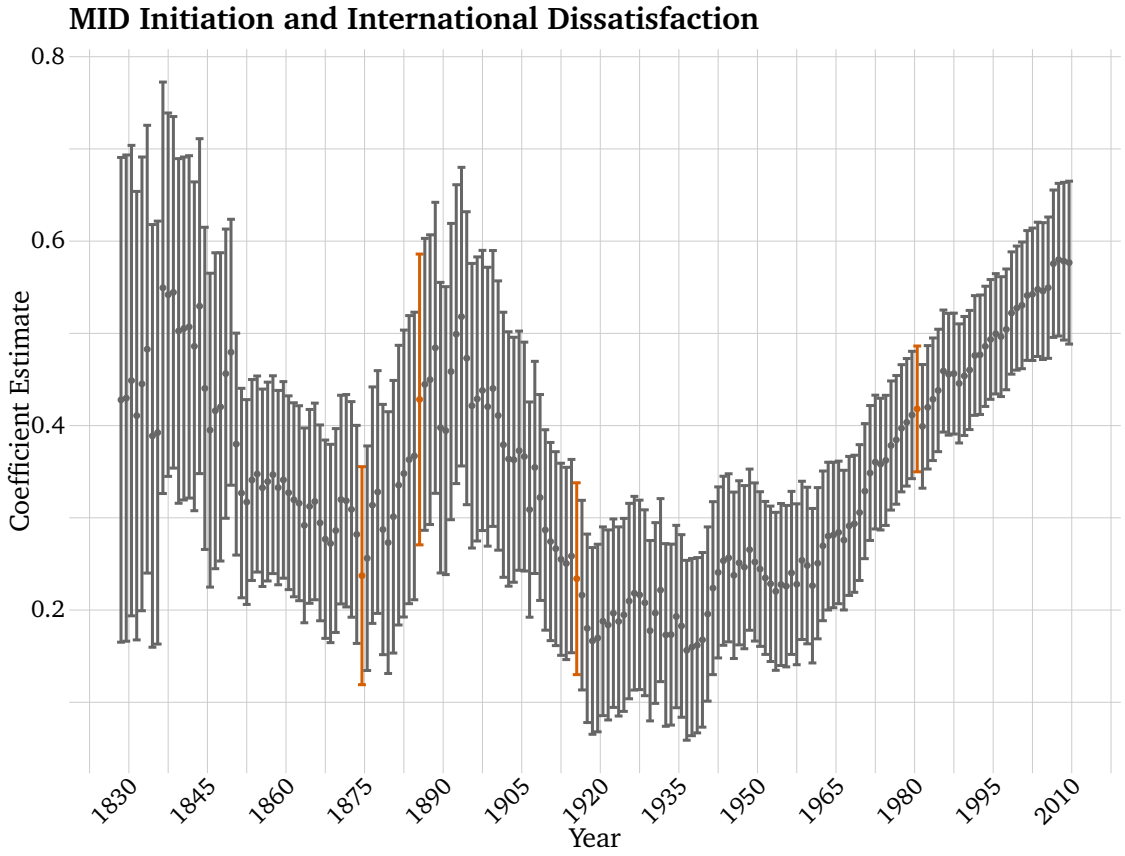

Coefficient estimates and standard errors for GLMs where the outcome is whether a state initiated a reciprocated MID in a country-year. Models with no control variables are a regression of conflict onset on a country's dissatisfaction. Models with control variables include a country's polity score, their ideal point estimate from UN voting records, and their population in a given year. Beyond wanting to compare the dissatisfaction measure to a country's ideal point, the other two control variables are included for two reasons. First, studies of conflict tend to consider regime type and size as control variables. Second, it is possible that both controls are causally prior to dissatisfaction and conflict. All models are fit with and without fixed effects. Unsurprisingly, coefficients are pushed closer to zero with fixed effects models. Count models of the number of MIDs initiated in a country-year are also estimated.

Figure 1.7 includes the resulting coefficient estimates. While effect magnitude depends

upon whether fixed effects and control variables are included, the primary takeaway point is that international dissatisfaction is, regardless of model specifications, significantly associated with a positive increase in the probability of conflict initiation. Insofar as predicted probabilities are concerned, comparing predictive performance and case-specific predictions is the focus of my next chapter. However, in this regression-specific context on modeling linear associations, when fixed effects are not included, then countries that find themselves further on the fringe of ideal point estimates are more likely to also initiate MIDs. *Yet*, when fixed effects are added to the model, then the relationship between ideal points and MID onset completely reverses and becomes negative. This model-dependence suggests that ideal point estimates are useful in a dyadic context, speaking to how different any two states are. But if one's goal is to understand any one state's general dissatisfaction with the status quo, then the measure provided here provides more reliable statistical properties.

Moreover, the relationship between international dissatisfaction and conflict initiation is not just resilient to control variables and the outcome's distribution. Indeed, the relationship's direction and significance is also time-invariant. Figure 1.8, includes the results of a vary-coefficient model (Hastie and Tibshirani, 1993), which are then tested for locations in time where the underlying coefficient changes – rather than just moving due to random noise. While some changepoints are detected, meaning the relationship's magnitude does vary (Kent et al., 2020), the general point about international dissatisfaction likely driving conflict onset across time and space holds, with the estimated coefficients always being positively associated with onset. That said, that the relationship does vary over time presents substantively important and interesting variation, which itself is worth investigating. Subsequent work should ask: Why does the relationship between dissatisfaction and conflict onset vary when it does? And why the consistent increase from the end of World War II forward?

Figure 1.8: GLM Coefficients Over Time




Regression coefficients for a logistic regression of MID initiation in a country-year on international dissatisfaction, controlling for ideal points, polity scores, and the logarithm of a country's total population. Notably, all coefficients are greater than zero and statistically significant. That being said, coefficient magnitude does vary in size. Coefficients which are estimated to be changepoints, based on effect magnitude, are highlighted in red and larger. Whether the relationship between international dissatisfaction and conflict is time-variant and why is an interesting question for future work.

## 1.6 Implications

Continuing with the questions raised by coefficient changepoints, the measure's statistical performance not only raises multiple substantive implications, but also novel research questions. Substantively, the measure captures whether or not a country has reason to think that its access to valued international goods falls short of what it should be able to gain in

a coercive bargaining scenario. Importantly, the measure represents a form of social exclusion, where a country compares the sum of its relations relative to its overall capabilities. In this sense, small states may generally be low in international benefits, but also low in expectations about what they can realistically achieve. Furthermore, because expectations are low, achieving equilibrium is also relatively likely. Joining on to a larger power's hierarchy – exchanging autonomy for welfare (Lake, 2009) – provides ample access to international goods and services.

Rather, by the measure's logic, the greatest threat comes from countries that are large enough to have substantial expectations, but find themselves relatively excluded from the necessary communities and networks that will satisfy unmet expectations. In many ways, this is a fundamental characteristic of revisionist states throughout history. Whether it be Pre-WWI and WWII Germany, Imperial Japan, Communist Russia, or others, a hunger for unmet systemic change drove campaigns of international belligerence. While the story of revisionist states is unlikely to be monocausal – with robust literatures on the role of differential growth rates and various domestic factors demonstrating their importance – the measure's predictive performance does highlight general dissatisfaction's centrality to belligerent states.

Expanding on the point of international revision likely resulting from multiple processes – not just dissatisfaction via social exclusion – a comparison to other explanations for revision is merited. The results in this paper suggest that measuring international dissatisfaction in a Gilpinian manner captures conflictual tendencies more than fringe ideal points, which is a useful validation of the measure. But does the measure itself better predict conflictual tendencies than other popular classes of explanation for international revision? More specifically, does international dissatisfaction better predict a state's propensity to start conflicts than differential growth rates or drastic domestic changes? While a different question than why states become dissatisfied in the first place, knowing which class of variables best predicts conflictual tendencies is valuable for policy and scholarship alike, as it can provide analytical focus. In the next chapter I take on this question, comparing the predictive ca-

pacity of each variable in identical machine learning formats with identical training and test sets.

Before concluding, the evidence found in this paper carries provocative and novel implications for debates around whether China is or will be a revisionist state. (Johnston, 2003) *If dissatisfaction stems from a state's ability to access valued international goods, then whether or not China pursues a strategy of revision may depend more on the rest of the international system than China itself.* Put differently, if potential partners are willing to deepen relations with China as its capabilities grow, then there will likely be no incentive for China to alter a system which continues to match its rise. *But if China's growth continues to outpace other state's willingness to expand relations with it, then we should not be surprised by a substantial uptick in the coercive use of Chinese military force.* The paper's results – and those in the next chapter – call for a focus on the relationships between China and its peers and asking whether we should expect those relations to grow stronger or to fray in the coming years.

## 1.7 Conclusion

International dissatisfaction is widely recognized as central to theories of international conflict and cooperation. However, despites its theoretical significance, its unobservability presents a barrier to empirical operationalization. In this paper I take Gilpin's framework for international dissatisfaction and produce statistical approximations. The result is a dataset with estimates for every state's international dissatisfaction from 1816-2012. When compared to a potential alternative – estimates for a state's ideal point based on its voting record at the United Nations – the measure is a far more robust predictor of conflict onset. Moreover, when estimated with a varying-coefficient model the relationship between international dissatisfaction and conflict onset remains positive and statistically significant across all available years. While dissatisfaction – measured via relative social exclusion – likely does not alone explain revisionist tendencies, the robust relationship between dissatisfaction and conflict onset does support dissatisfaction's centrality to international revision.

The measure also provides research opportunities for any area where dissatisfaction plays a conceptual.

# Chapter 2: Predicting International Conflict and Revision

## 2.1 Introduction

Political decision-makers face a choice at all times between pushing their country in a direction that supports or seeks to disrupt the international status quo. This fundamental choice between international revision and status quo seeking is widely understood to be highly consequential.[1] The world would be vastly different if Germany had been satisfied with the European order before either of the World Wars or if Revolutionary France did not turn its guns outward. In this sense, the typology of status quo and revisionist states is not just a conceptual exercise used to simplify a complicated world; it is a categorization that captures one of the defining variables in international politics. Whether it be Napoleonic France, Revolutionary Iran, Communist Russia, Imperial Japan, or 20th-century Germany, revisionist states have underscored the outbreak of some of history's bloodiest wars and costliest military competitions.

Indeed, one of the most discussed questions in modern international politics is whether China's long-term aims are likely to follow a revisionist or status quo path. (E.g. Christensen, 2006; Goh, 2013; Johnston and Ross, 1999; Johnston, 2003; Paul, 2016; Schweller, 1999) If China's strategic trajectory involves upending the current international order, then

[1]The impact of this choice can be seen in the lengthy literature discussing the role of status-quo and revisionist states in international politics. (E.g., Carr (1946, 103-105), Morgenthau (1948, 40-74), Wolfers (1962, 81-102), Organski and Kugler (1981), Schweller (1994), Lyall (2005), and Goddard (2018b))

the prescriptions that follow diverge widely from those toward a China that intends to maintain the status quo which enabled its rise. Accurate predictions about which path China is most likely follow under current and alternative conditions are therefore a policy necessity. So when do countries – particularly revisionists – fall where they do on this spectrum? Put differently, when do countries sometimes attempt to forcefully undermine or alter the international system, rather than invest in and support it?

In this paper I present an empirical comparison of the three most prominent arguments about revision's origins. Generally speaking, arguments fall under one of three themes: 1) rising powers and differential growth rates, 2) domestic political shifts and revolutionary regimes, or 3) international dissatisfaction. Each of these arguments is backed by various historical cases and is statistically associated with interstate conflict initiation. However, their relative ability to inform accurate policy predictions is an open question. Put differently, *it is unclear which theory provides the most predictive accuracy*, even though we have reason to expect that all do predict to some extent. Or, if we know about a state's domestic political changes, growth rates, and international dissatisfaction, then how much of a state's revisionist behavior can we predict with each variable and which variable informs the most accurate predictions?

Herein I compare the three theories empirically, building stacked ensembles of machine learning models that vary only in the features included for prediction. The target (outcome or dependent variable) is whether or not a state initiates a militarized interstate dispute in a given year. In order to compare each theory empirically, I measure and compare each ensemble's predictive accuracy in test sets which did not inform model training. Across almost all time periods, the ensemble based on international dissatisfaction predicts with the highest accuracy in test sets whose data did not inform model training. Notably, the other two ensembles also predict with reasonable accuracy in test sets, meaning the variables associated with domestic political changes and rising powers do provide meaningful, even if less, predictive capacity of a state's revisionist tendencies.

The remainder of the paper proceeds as follows. First, I overview the literature on

revisionist states, discussing the three aforementioned classes of arguments in detail and setting up their empirical operationalizations. Next, I discuss how features were engineered from the available data and the methodological approach behind the machine learning models employed, emphasizing that this is a predictive, rather than a causal approach. This is followed by a presentation of model results, with an emphasis on moving from a "black-box" view of machine learning models to one where results can be interpreted in a substantively useful way, speaking to the theories that inform the variables chosen for each model. (Karpatne et al., 2017; Radford and Joseph, 2020) I close with a discussion of implications that follow from the paper's results regarding China and other great powers in the current era. Given the paper's results, I also propose potential avenues for future research – emphasizing an understanding of why some states find themselves relatively excluded from valued international institutions and communities while others are able to deeply embed themselves with general ease. The paper's predictive focus leaves it agnostic on the causal question of what leads to exclusion and the resulting dissatisfaction. However, dissatisfaction's predictive accuracy when it comes to interstate conflict emphasizes the need for future work to investigate why dissatisfaction occurs.

## 2.2 Theories of International Revision and Conflict

Prominent arguments about revisionist states tend to fall under one of three categories: rising powers, domestic political shocks, and international dissatisfaction. In other words, discussions of revision tend to stem from, respectively, the threat of a rising power, states with a drastically changing domestic political landscape, or an international system whose characteristics are viewed as untenable by certain key members. While these themes are investigated and argued to impact states in numerous ways, few, if any, arguments regarding international revision fall outside of the three. Below, I start with the logic of rising powers and international revision before turning to discussions of domestic politics and then closing with international dissatisfaction.

Starting with Thucydides[2] and generally discussed in terms of Organski and Kugler (1981), power transition theory links international revision to rising powers who threaten using their newfound or forthcoming capabilities to forcefully reshape the international system.[3] As a rising power grows stronger it threatens to overtake an established power's position as the systemic or regional hegemon. A commitment problem is found at the heart of power transition theory and is the reason why power transitions are argued to produce war (Fearon, 1995). Rising powers cannot credibly commit to using their new capabilities in a future manner that is acceptable for the established powers, because the state's future intentions and leadership are largely unknown, especially as one's strategic time horizon's extend further into the future (Edelstein, 2017; Tingley, 2011). Even if relations with the rising power are currently healthy for many states, the future is largely unknown and could very plausibly change radically to more conflictual relationships.

When this uncertainty about future intentions and a state's inability to commit indefinitely to any course of action are coupled with an impending power transition, then states find themselves in a uniquely precarious situation. Established powers may make a strategic calculus that starting a war under current conditions – where they are still more powerful than the rising power – is considered a better option than living in future subservience after a transition, despite the unavoidable costs in blood and treasure that come with war. Accordingly, a war-avoiding bargain between established and rising powers cannot be reached. Moreover, while this process is generally discussed in the context of power transitions – where the stakes of the commitment problem are most severe – the logic of rising powers applies to any situation where one state's growth rate is greater than its actual or potential competitors' growth rates. Even without a power-transition, if a state is growing rapidly and expects to continue growing, then it can use its newfound capabilities to reshape its environment in a way that puts it in a position of taking further advantage of other states in

[2]See the recently re-popularized discussion of the "Thucydides Trap" in Allison (2017), where the core claim being applied to the US-China relationship is: "It was the rise of Athens, and the fear that this instilled in Sparta, that made war inevitable."

[3]Additional treatments include, but are not limited to: Christensen (2006); Duffy Toft (2007); Lebow and Valentino (2009); Kennedy (2010); Mearsheimer (2014); Dafoe et al. (2014).

the future. Subsequent work has delved into the conditions under which these commitment problems are more or less severe, with an emphasis on: the number of other competitor states to consider (Shifrinson, 2018; Snidal, 1991), the compatibility of interests between actors (Schake, 2017), the strategic time horizons of key leadership (Edelstein, 2017; Tingley, 2011), and the relative plausibility of an actual power transition (Beckley, 2012, 2018; Kadera, 2001).

Secondarily, much of the relevant literature focuses on the domestic actors who decide to pursue a strategy of international revision. Revisionist policies are debated, decided upon, and implemented from within a country. So it follows to ask why decision-makers within certain domestic political systems and regimes may be more likely than their counterparts in other countries to pursue international revision. On this point, many famous revisionist states are inescapably linked to their unique domestic political environment at the time and are rarely discussed without mention of the domestic politics and leadership at the time of revision. Napoleonic France, Hitlerite Germany, Revolutionary Iran, Imperial Japan, and Communist Russia are core examples of revisionist states and almost always referred to in the context of these domestic labels. For these examples, their revisionist behavior can be viewed as intertwined with the radical domestic changes that preceded a decision to attempt widespread revision of their international environments. Under a counterfactual hypothetical where the international environment is the same but domestic leadership differs for these countries, revision appears far less likely.

Beyond the correlation between revision and major domestic changes in these historical cases, conceptually it is an appealing argument that political elites who are seeking radical political change will not just stop at home, but see similar grievances and opportunities abroad. Once major political changes have been profitable for instrumental elite actors at home, the next step may very well be to look for opportunities for further similar changes abroad. However, in response, outside actors may also see radical domestic changes within their neighboring states and grow fearful, creating an increasingly tense international environment which is ripe to spiral into interstate conflict. (Walt, 1996) Taken as a whole,

whether it be the rise of militaristic regimes and leadership[4], the prominence of overexpansive and snowballing political projects,[5], or domestic revolutions and civil wars turning their focus abroad[6], various mechanisms have been theorized to link sharp domestic political changes to subsequent international revision.[7]

Lastly, by definition, for revision to occur, there must be an environment to revise. The final class of arguments focuses on the structural makeup of each state's international environment. Here, the question is: what structural conditions are most prone to push states toward attempting to revise the international system for their benefit, despite the massive risks? These arguments can be generally cast as discussions of international dissatisfaction, where a state pursues revision because it views some aspect(s) of the international system as stifling and unacceptable. Notably, this class of arguments is amenable to both rationalist and more constructivist or sociological explanations. Across the literature, approaches span from identity (Thies and Nieman, 2017), linguistic (Goddard, 2018a), norms (Finnemore and Sikkink, 1998), relational (Jackson and Nexon, 1999; Qin, 2016; MacDonald, 2014), and status concerns[8] to more calculated cost-benefit analyses.[9] For this literature the greatest concern should be directed toward states with the capacity to coercively create widespread international change – whether or not they are concurrently a rising power – with longstanding international grievances and/or reasons to think that its environment can feasibly be reshaped in a way that better suits its interests. In other words, large dissatisfied states are the most dangerous and likely candidates for deciding they want to pursue a strategy of widespread international revision and conflict.

Across all of the aforementioned theories and arguments, statistically significant associ-

[4]Horowitz et al. (2015); Lemke and Reed (1996); Schweller (1994, 1999, 2015); Snyder (1984); Van Evera (1984); Weeks (2008)

[5]Davidson (2006); Lyall (2005); Snyder (1991)

[6]Lawson (2015); Schroeder (1994); Walt (1996)

[7]On the China debate, given the prominence of the CCP and Xi Jinping's leadership in political discussions, domestic theories of international conflict and revision are potentially of primary importance for current policy options.

[8]Chan (2004); Deng (2008); Duque (2018); Goh (2013); Larson and Shevchenko (2010); Paul et al. (2014); Renshon (2016, 2017); Ward (2017); Wolf (2011)

[9]Braumoeller (2013); Holsti et al. (2019); Gilpin (1983); Goddard (2018b); MacDonald and Parent (2018); Montgomery (2016); Morgenthau (1948); Carr (1946); Wolfers (1962); Trachtenberg (2012); Lipscy (2017)

ations and detailed case-studies are abundant. In this sense, all variables of interest likely matter and are present in seminal cases to some degree. However, saying that many variables have some importance and can predict revisionist states is not the same as saying their predictive capacities are of equal magnitude. While that may be the case, it seems more likely that the mechanisms precede revision at differential rates. Furthermore, when it comes to projections and policy prescriptions about current and future potential revisionists, knowing which theoretical pathway has the greatest empirical support is vital. In the following sections I provide an empirical composition and comparison of each. After presenting empirical operationalizations of the general arguments, I evaluate each theory by comparing predictive accuracy across identical statistical frameworks and data.

## 2.3 Data

The outcome of interest for this study is whether or not a country initiates a reciprocated militarized interstate dispute (MID) in a given year. I treat whether or not a state initiates a MID in a given year as the dependent variable because a defining characteristic of revisionist states is that they are especially conflict-prone, relative to other states. This is not to say that all interstate conflicts are revisionist in nature,[10] but models meant to capture the underlying causes of revision should predict conflict onset well. Indeed, if variables meant to captures revision's origins do not accurately predict interstate conflict when modelled statistically, then this would put arguments that those variables are behind most cases of revision on shaky footing. Moreover, in the subsequent section with model results, when a model is fit with all possible features, controlling for other sources of conflict, international dissatisfaction is given primary importance by the model – generally providing the greatest increase in predictive accuracy across algorithmic iterations. Accordingly, my focus is on finding which variable both best predicts conflict onset when modelled on its own *and*

[10] On this note, the subsequent models only predicting a portion of all MIDs, which we should expect because the following models are not meant to be the models of conflict. Rather, we are asking: For the variables that are understood to often precede revision, which best predicts conflict?

alongside other variables also theorized to precede international revision. A variable which best predicts interstate conflict both on its own and alongside other theoretically-informed variables is likely going to best predict the emergence of revisionist states.[11]

However, I do not use all MIDs as my dependent variable. Rather, per Braumoeller (2019), employing reciprocated MIDs strikes a useful balance between modeling all MIDs and only looking at cases which resulted in a fatality. Because the MID dataset (Palmer et al., 2019) includes cases that range from low-level threats of force to interstate wars, it is easy to lump together non-threatening posturing with intense outbreaks of military conflicts. While fatal MIDs are likely too strict of a cutoff, eliminating cases that could have easily escalated further but fortunately did not, using all MIDs is likely too lenient of a standard, including low-level incidents such as fishing trawlers entering into territorial waters.[12] Instead, as Braumoeller (2019) points out, if the use of force is reciprocated, then both sides have demonstrated a willingness and ability to escalate.[13] In terms of revisionist states, as discussed above, while not all uses of force are revisionist in nature, revisionist states are especially prone to military action, so looking at a state's propensity to initiate reciprocated MIDs is an effective proxy metric. Looking at the data, of the 15925 total country-years, 3167 include the onset of a reciprocated MID. This leaves us with the common problem of class imbalance, where the dependent variable is mostly zeros – risking models that predict almost entirely zeros and returning often accurate but generally unhelpful predictions. While the following models are able to make accurate predictions without techniques such as downsampling or upsampling, I ultimately turn to downsampling because it provides the highest predictive accuracy across the entire training set, cross-validation folds, and test set.

Turning to features, each of the aforementioned theories is represented by a core vari-

[11] I make this claim recognizing that interstate conflicts also occur for a variety of non-revisionist reasons. Seeing dissatisfaction's subsequent predictive capacity for conflict, despite including an array of other features for prediction, provides greater confidence in the results speaking directly to international revision.

[12] That said, Figure 9 in the appendix includes model results when fatal MIDs are used as the dependent variable and the results are consistent with reciprocated MIDs. Also in Figure 10 I show model results when the DV is made even more fine-grained and limited to wars – or MIDs with 1000+ casualties. For these models, dissatisfaction also demonstrates the generally strongest predictive capacity.

[13] Moreover, as Braumoeller (2019) also points out, interstate conflicts can escalate very quickly in dramatic fashion. So just because a reciprocated MID did not escalate beyond low-levels of force does not, by any means, suggest that further escalation was impossible.

able, which are then each transformed in a set of identical ways. The three core variables are a state's international dissatisfaction, expected international benefits, and its domestic political regime. The latter two variables are then first-differenced in order to represent rising powers and domestic political changes.[14] As I discuss subsequently, estimates for a state's international expectations are based on its observable capabilities and general reputation for effectively using its capabilities, both of which then inform how much a state expects it should be able to influence and benefit from the international system. Taking the first difference of these expectations therefore captures whether a state is rising, declining, or staying at its current size. A state's international dissatisfaction in a given year is operationalized as the difference between its expected international benefits and actual international benefits. The more a state's actual access to valued international goods and social recognition falls short of what it believes it should be receiving based on its relative capabilities, then the more that state will be dissatisfied with the status quo. The more dissatisfied a state is, then the more we expect it to start interstate conflicts. Put generally, dissatisfaction for state $i$ in year $t$ is expressed as:

$$\text{Dissatisfaction}_{it} = \log(\text{Expected Benefits})_{it} - \log(\text{Actual Benefits})_{it}. \tag{2.1}$$

Both expected and actual benefits are logged because the variables – which I subsequently break down – are heavily right skewed, so logging them provides us with approximately normal distributions, avoiding a situation where the handful of extreme outliers at the untransformed distribution's tail drive inferences disproportionately.[15] Both a state's expected and actual benefits are variables that I construct and require feature engineering with multiple readily available datasets.

Starting with the measure of actual international benefits, the actual benefits for state $i$ in year $t$ can be expressed as:

[14] These variables all cover very different processes. Figure 2.1 displays the distributions of each component variable and their correlations, where the strongest correlation between any two features is below 0.5.

[15] This of course builds in an assumption that working with normally distributed data is preferable to skewed data and either makes little difference to or improves our final substantive interpretations.

$$\text{Actual Benefits}_{it} = \frac{1}{N}\sum_{n=1}^{N}\text{PageRank}_{int} \tag{2.2}$$

with $\text{PageRank}_{int}$ referring to state $i$'s PageRank centrality in year $t$ in a network $n$ (including military alliances, interstate trade, shared diplomatic relations, and arms transfers), which is then averaged to provide a state's general access to valued international goods in a given year. The idea behind using network centrality to measure actual benefits is that prominence in these valued networks provides access to the social and material goods that constitute the potential profits from international politics. The PageRank centrality formula (Brin and Page, 1998, 2012) estimates a node's prominence in a network, which reflects the sum of a node's ties, with ties to other prominent nodes weighted more heavily than ties to less prominent nodes.[16]

$$x_i = \alpha_A \sum_j a_{ij}\frac{x_j}{g_j} + (1-\alpha_A)\frac{1}{N} \tag{2.3}$$

where $g_j = \sum_i a_{ji}$ (node $j$'s total number of ties), $x_i$ is the PageRank centrality for node $i$, and $x_j$ is the PageRank centrality for node $j$.[17] $a_{ij}$ is equal to 1 if a tie exists between nodes $i$ and $j$, but it is equal to 0 otherwise. $\alpha_A$ is a constant damping factor that weighs how much a node's ties matter, as opposed to treating each node as equally central (if $\alpha_A = 1$); in the context of Google's search engine the damping factor approximates the probability that a user will stop clicking links at any time.

A state's expected benefits are produced by multiplying its CINC score (the state's share of the entire globe's observable material capabilities) by its average predicted probability of winning a MID against other states (a proxy for the state's perceived capabilities):

$$\text{Expected Benefits}_{it} = \text{CINC}_{it} \times \frac{\sum_{j=1}^{J}\Pr(\text{Win}_j)}{J-1}. \tag{2.4}$$

[16]The original idea was that the internet can be viewed as a network where nodes are sites and ties are links to one site that are included on another site. Google's search engine then ranks results based upon a version of PageRank centrality, where a site's ranking on the search algorithm is based upon its PageRank centrality.

[17]If one carefully reads the formula, then they will see that calculating a node's PageRank centrality requires having already estimated every other node's PageRank centrality. This produces a chicken-egg problem where there is no obvious answer about how to estimate initial values. For a mathematically detailed explanation of how this problem is overcome see: http://www.ams.org/publicoutreach/feature-column/fcarc-pagerank.

While CINC scores captures a state's percentage of the globe's material capabilities – such as iron and steel production and total population – the reputational sources of power are much more difficult to model. Per Gilpin (1983), I decompose power into observed material capabilities and a state's reputation for using those capabilities. In order to model the latter, I build on Carroll and Kenkel (2016) and train a machine learning model on MID outcomes, building a model that predicts which state wins interstate conflicts based on every involved state's observed CINC components. After accurately classifying observed MID outcomes, I then apply the model to make predictions about which state would win in all possible MIDs. A state's reputation for power – $\frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$ – is then the sum of a state $j$'s predicted probabilities of winning across all possible MIDs, divided by the number of other states in the international system in a given year.

Combining these two variables produces a measure of expected international benefits which weighs a state's observed capabilities by its reputation for the capacity and willingness to use its capabilities. This combined measure represents how powerful a state is, relative to its peers, which then drives how much that state expects it should be receiving from the international system at any time. The more powerful a state is, then the more it should expect the system will be designed in a way that benefits it. Inversely, the less powerful a state is, then the lower its expectations will be, because it lacks the coercive capacity to meaningfully shape its international environment. Taken in comparison to its actual international benefits, the difference between a state's expected and actual international benefits therefore represents the extent to which a state believes the international system is currently depriving it of benefits that it could otherwise gain through a coercive bargaining scenario. This expectations-reality gap then creates an incentive for using military force to remedy these perceived systemic shortcomings, which otherwise appear to be persistent.

Moreover, because a state's expected international benefits represent its relative power – or share of global capabilities – the measure maps on cleanly to theories of rising powers. Because rising powers occur when a state is growing faster than its peers, if a state's share

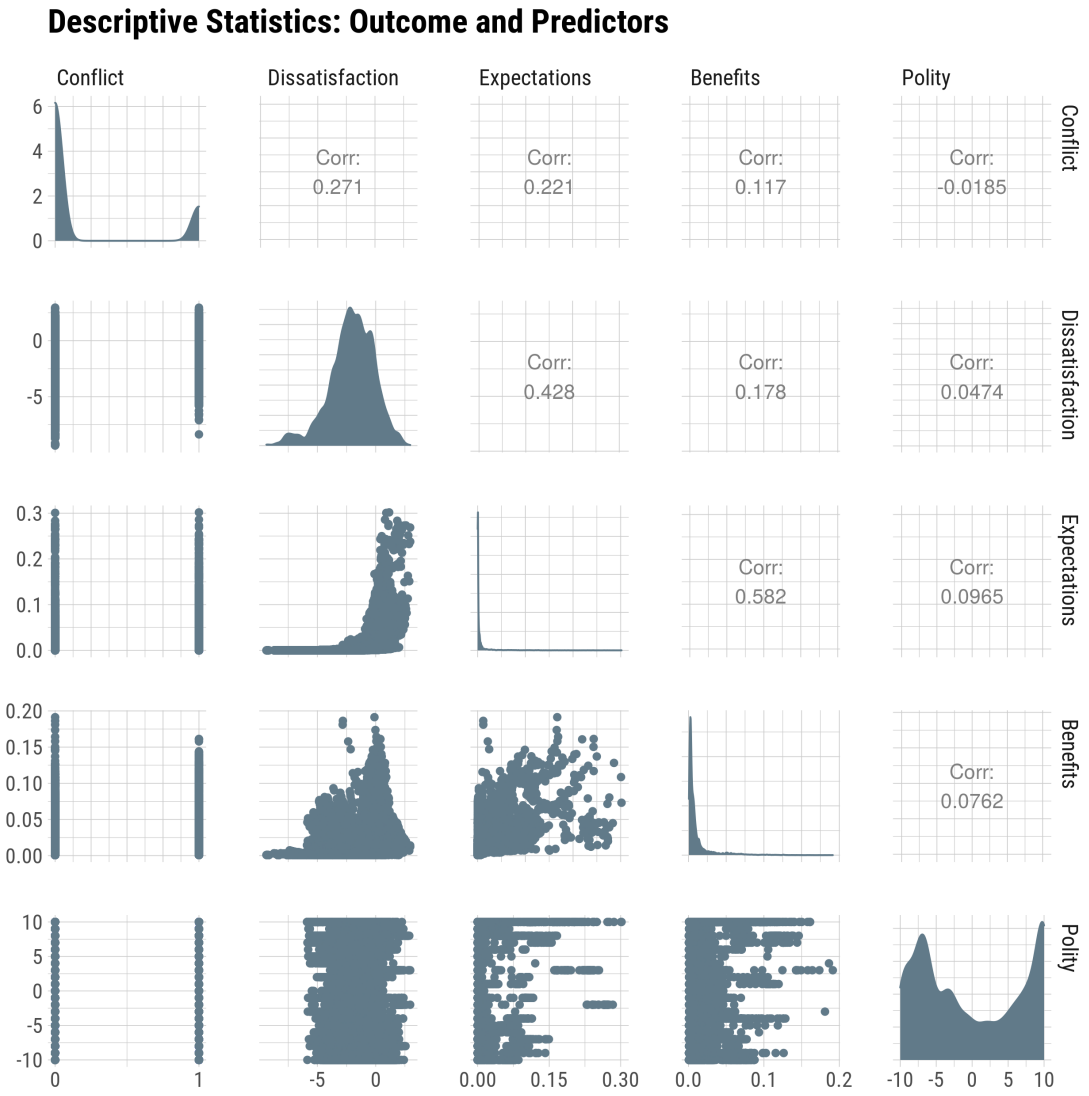
of global capabilities grows, then it by definition is growing faster than its peers and is a rising power. Accordingly, our measure of expected benefits is well-suited as a starting point for operationalizing whether or not a state is rising and the extent of its differential growth rate. More formally, we can capture differential growth rates by first-differencing a state's expected benefits – calculating the difference between a state's expected benefits in year $t$ and year $t-1$. If the difference is positive, then the state is rising by that extent. Admittedly, differential growth rates are of most concern for international revision when the great powers are growing and this empirical measure only compares growth at time $t$ and $t-1$, regardless of a state's initial size. While this is conceptually a concern, in terms of this measure – where a state's size is represented in terms of its share of the entire international system's capabilities – large growth rates only tend to occur within states that are already larger than most and have the capacity to expand relative to other states. This point is demonstrated in Figure 10 in the appendix. Returning to operationalizing growth rates, if the difference is negative, then that state is shrinking relative to its peers. While rising powers are often discussed as a binary category, this approach gives us a finer-grained continuous measure of differential growth rates:

$$\text{Differential Growth}_{it} = \text{Expected Benefits}_{it} - \text{Expected Benefits}_{it-1}. \tag{2.5}$$

Lastly, I consider three variables for relevant theories of domestic politics and international revision, where the primary variable is a state's polity score (Marshall et al., 2002), because of its conceptual utility and availability for all states in all years. Polity scores are used to capture regime type and span from fully autocratic (-10) to fully democratic (10). Similar to rising powers, domestic theories of revision are generally linked to drastic changes to a state's domestic political environment, so like rising powers a use a first-differencing strategy to capture changes in a state's regime type from one year to the next. However, adding some conceptual richness, drawing on theories of revolution and war, I include variables for whether or not a civil war began or if a civil war is ongoing in a country in a given year. While all of our other data starts in 1816, data on civil wars starts in 1945

and is subsequently only considered in models military conflict after World War II.

Figure 2.1: Distributions and Correlations Between Outcome and Features

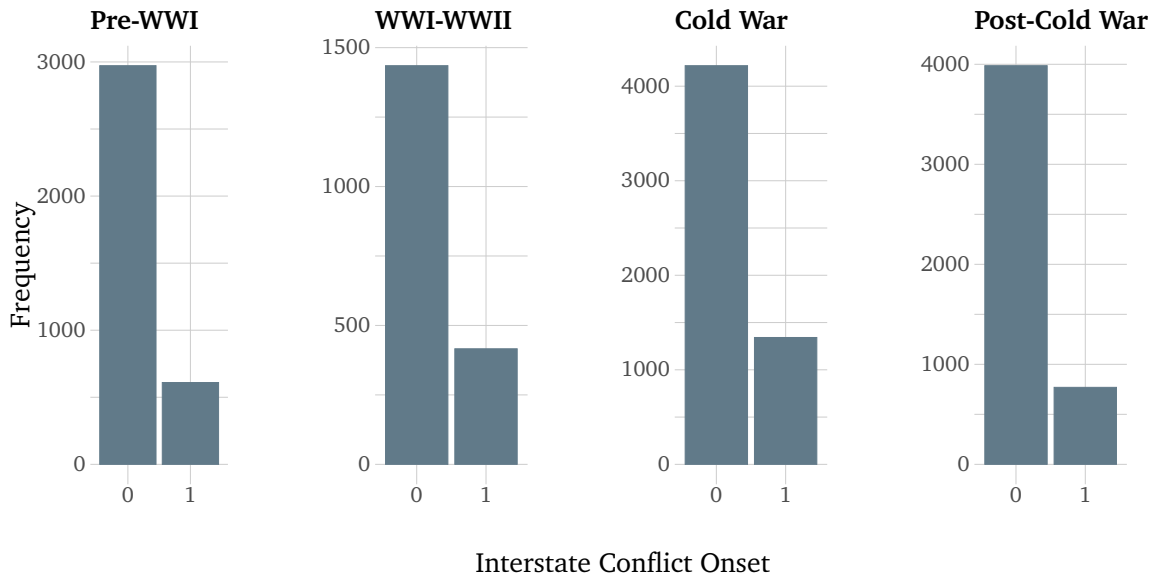



Distribution of and correlation between outcome and core features. Diagonal elements are the distribution for each variable. Elements to the left of the diagonal are scatterplots of each variable against another. Elements to the right of the diagonal list the correlation between each variable.

In terms of our dependent variable, class imbalance (Japkowicz and Stephen, 2002) is a concern, where most country-years do not include any interstate conflict onset. This is

a common concern across many domains, where highly consequential events are relatively rare. In situations of class-imbalance a model with high accuracy is likely to incorrectly classify the minority class (here interstate conflict onset), which actually has more substantive importance. This data is no different, though the imbalance is not as intense as if this were a dyadic study of conflict. The imbalance is visualized in Figure 2.2, with conflict onset being unbalanced not just in the aggregate, but across all time periods. In the following section I discuss sampling techniques used to create balance during model training.[18]

Figure 2.2: Outcome Distribution Across Time Periods




Outcome distribution across time periods. For each major era of international history interstate conflict onset is characterized by class imbalance, where most country-years are characterized by the absence of interstate conflict.

[18]Sampling methods – such as upsampling, downsampling, or SMOTE – are not the only way to deal with class imbalance, with zero-inflated models, and additional data gathering as alternative approaches. In our case, we have exhausted all available data and sampling methods proved more effective than zero-inflated regression models for this data.

## 2.4 Methods

### 2.4.1 Comparing Predictive Accuracy

Rather than comparing regression coefficients or pursuing causal identification, I approach this data from a *predictive* perspective (Cranmer and Desmarais, 2017; Shmueli, 2010). In a traditional regression format, producing statistically significant coefficients for this data is a relatively straightforward task. However, testing for non-zero coefficients is not always equivalent to evaluating which variables have the strongest empirical link with conflict onset, especially when variables are of different scales and relationships are likely complex and non-linear. Instead, for each variable of interest I apply identical predictive models, feature engineering, cross-validation, and training and test sets. This leaves every model with identical modeling procedures but different predictors. Each variable is then evaluated through its predictive accuracy on identical test set data. While the purpose of this modeling approach is straightforward, in practice it presents an especially difficult standard. As one popular Bayesian statistics textbook puts it: "Fitting is easy; prediction is hard."[19]

To set up a predictive architecture, I estimate a series of machine learning models, predicting whether or not a country starts a reciprocated militarized interstate dispute in a given year. These initial models serve as "base learners" whose predictions are then run through a stacked ensemble, or "metalearner", in order to produce a final set of predictions. Each set of base learners is identical in format and data, varying only in the included predictors. The resulting product is three sets of predictions, each resulting from variables representative of a different theory. More formally, each theory is represented using identical feature engineering applied to a single starting variable – differential growth rates, yearly changes in a country's regime type, or international dissatisfaction. For each of these predictors, let $x_{it}$ represent a variable's value for country $i$ in year $t$, the general model is:

[19] (McElreath, 2020, p12)

$$y_{it} = f\left(x_{it},\ x_{it-1},\ x_{it-2},\ (x_{it} - x_{it-1}),\ (x_{it} - x_{it-2}),\ x_{it}^2,\ x_{it}^3\right) + \varepsilon_{it}. \tag{2.6}$$

Each theory is therefore represented by a core variable, which is: lagged by 1 year, lagged by 2 years, first-differenced, second-differenced, squared, and cubed.[20]

As I discuss next, this predictive architecture is not set up to develop a new method or produce regression coefficients that can be tested in the null-hypothesis framework. Rather, the goal is to build a set of flexible machine learning models that draw on cutting-edge techniques from the statistical learning literature, allowing for accurate predictions of when interstate conflicts occur and investigating which variables contribute the most to accurate predictions.

### 2.4.2 How This Approach Differs

In most Political Science papers, statistical results are communicated by fitting a generalized linear model to the entirety of one's data and presenting coefficient estimates in a table or dot-whisker plot. The approach outlined above differs from this practice in three important ways. First, while the traditional approach in Political Science evaluates variables by the statistical significance and direction of regression coefficients for the *entire dataset*, I instead measure a variable's statistical contributions by its predictive accuracy in a *test-set* – data which the model did not interact with when being fit. Model evaluation based on test-set predictive accuracy guards against inferring statistical quantities when one is actually overfitting and treating the case-specific nuances of various data points as if they were representative of the underlying characteristics of all data points. Moreover, if the model is able to accurately predict on data that did not contribute to training, then the model likely is accurately representing the true data-generating process.

[20]The only exception is models of domestic changes from 1945 to the present, where binary variables are included for whether a civil war has begun or is ongoing. On this note, I fit multiple ensembles for each major period of international history, evaluating if the relative predictive accuracies are time-varying. This is driven by the fact that recent research suggests that the data-generating process for wars is time-varying (Anderson et al., 2016; Braumoeller, 2019; Jenke and Gelpi, 2017).

Second, a particular benefit of the gradient boosting approach is that its training process increasingly weights observations that are difficult to classify. Each successive tree is fit to the last tree's residuals, meaning the goal is to predict observations which have previously been inaccurately modeled. This process is inherently well-suited for modeling rare events like interstate conflict because the first iteration of model-building will generally find that the highest-accuracy set of predictions is to just classify all cases as being peaceful. However, this leads to residuals being almost, if not entirely, cases of conflict onset, which the boosting machine will focus on more. With interstate conflict being notoriously difficult to predict and characterized by large class imbalance, this approach has a substantial advantage over straightforward generalized linear models, which treat all observations as equally important. Indeed, as Figure 2.5 demonstrates, the gradient boosting machine weighs a state's international dissatisfaction especially heavily in the training process. Considered in tandem with the gradient boosting machine's strengths for this data, this lends additional support for dissatisfaction, relative to the other variables of interest.

Third, both the random forest and gradient boosting machine are tree-based methods, which allows for non-linear relationships. Unlike a generalized linear model, where a specific linear form is assumed for a conditional expectation function and then parameters are estimated for that functional form, tree-based models are flexible and can accommodate many types of relationships.[21] This flexibility can admittedly lead to increasingly complex model formulations where increased fit comes with a tradeoff of decreased interpretability. However, as I demonstrate in the next section, techniques have been developed for gaining a general sense of which variables contribute the most to the final model and how predicted probabilities tend to vary across different values for one's predictor of interest.

[21] In terms of Breiman (2001b), one approach to statistics assumes a set data-generating process and estimates parameter values for that form, whereas the statistical learning approach treats the functional form for the data-generating process as unknown and to be estimated algorithmically from the available data.

### 2.4.3 Ensemble Component Models

In this section I present the methodology for a series of machine learning models (base learners) which are then run through a final stacked ensemble to classify interstate conflict initiation. In order to make my modeling choices as clear as possible, I go into a substantial level of detail for each technique. The three base learners are a regularized logistic regression, random forest, and gradient boosting machine.

First, the regularized regression is a LASSO, which estimates coefficients that minimize the negative log-likelihood for:[22]

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\mathcal{L}(\beta_0,\beta;\boldsymbol{y},\boldsymbol{X})+\lambda||\beta||_1\right\}. \tag{2.7}$$

Here, $\boldsymbol{y}$ is a N-length vector of outcomes, $\boldsymbol{X}$ is a $N \times p$ matrix of the predictors, and $\mathcal{L}$ is the log-likelihood function for a GLM (in our case a logistic regression). $\lambda||\beta||_1$ is referred to as the $\ell 1$ penalty, which is defined as $\lambda\sum_{j=1}^{p}|\beta_j|$, or the sum of the absolute value of all coefficients multiplied by a parameter, $\lambda$. Through a bias-variance tradeoff, increasing $\lambda$ biases coefficients toward 0 but also decreases the model's variance, which increases accuracy when a model is otherwise prone to overfit. If $\lambda = 0$, then the model is equivalent to a traditional regression and if $\lambda = \infty$, then all coefficients will equal 0. While powerful and interpretable, the model does assume a linear relationship.

To relax this linearity assumption I next fit two tree-based classifiers, the first of which is a random forest. Rather than analytically solving for a set of coefficients, a random forest (Breiman, 2001a) builds a series of regression trees, where each tree provides a set of decision rules for predicting outcomes. The goal of a regression tree is to produce a set of non-overlapping regions, $R_1, R_2, ..., R_J$, with every observation that falls within a region, $R_j$, receiving the same prediction. These regions are decided upon through a partitioning process, where the model repeatedly evaluates which data point, $s$, within the available predictors, $X_j$, splits the data in a way that minimizes a within-group loss function. Random

[22]This notation is drawn from p.32 of Hastie et al. (2015). Hastie et al. (2009) and James et al. (2013) also discuss the LASSO in an acessable manner.

forests then build an ensemble of $B$ regression trees – creating a forest – and average across each tree's predictions. The forest is built through a 'bagging' procedure (Breiman, 1996), where: bootstrapped samples are drawn (with replacement), a tree is built for each sample, and then the predictions for all trees are aggregated into a single final prediction. More formally:[23][24]

1. For $b = 1$ to $B$:
   (a) Sample some data with size $N$ from the training data, with replacement.
   (b) Fit a regression tree, $\hat{f}^b$, to the sampled data, where the subsequent steps are repeated until a pre-specified minimum node size, $n_{min}$, is reached.
      i. Randomly sample $m$ of the available $p$ features.
      ii. Find the split point that minimizes some loss function among the variables, $m$.
      iii. Split that node into two daughter nodes
2. Any tree $b$'s classification for data point $x$ is labelled $\hat{C}^b(x)$.
3. The random forest classifies data point $x$ by evaluating:

$$\hat{C}^B_{rf}(x) = majority\ vote\{\hat{C}^b(x)\}^B_1 \tag{2.8}$$

   where *majority vote* refers to the most frequent classification across all trees.

In a gradient boosting machine, rather than fitting a multitude of trees at once, trees are fit consecutively with the goal of predicting the error term from the past model. By repeatedly fitting a model to the previous residuals, misclassified observations receive additional focus. Indeed, the algorithm is designed so that the residuals for each model represent the difference between $y_i$ and the sum of all previous models' predictions. Accordingly, once a tree correctly predicts the current residual, $r_i$, then the sum of the predictions across all models adds to $y_i$. Gradient boosting machines are powerful and generally more accurate

[23] See p588 and chapter 15 of Hastie et al. (2009) for this notation and a more detailed walkthrough.
[24] Generally for step ii, $m = \sqrt{p}$ or $m = \log_2 p$.

than random forests. However, the process of iteratively converging on $y_i$ also makes them prone to overfitting when $y_i$ is widely dispersed around the target function $F(x)$. In order to reduce overfitting, two parameters are generally applied. First, $\lambda$, determines the learning rate, or how much each model's predictions are included. Second, like in a random forest, $d$ determines how many splits are allowed for each tree – with fewer splits limiting the capacity of any one tree to overfit. Per James et al. (2013), the algorithm proceeds as follows:[25]

1. Set initial predictions as $f(x) = 0$ and initial residuals as $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1, 2, ..., B$, repeat:
   (a) Fit a regression tree, $\hat{f}^b$ with $d$ splits ($d+1$ terminal nodes) to the training data $(X, r)$.
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{2.9}$$
   where $\lambda$ is a regularization parameter, minimizing the risk of overfitting.
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) \tag{2.10}$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{2.11}$$

[25] p323

### 2.4.4 Stacked Ensemble

Once each base learner is fit, I run them all through a stacked ensemble, or 'super learner'.[26] The goal of a stacked ensemble is to combine a collection of models and make predictions based on each model's strengths (Hastie et al., 2009).[27] This occurs by inserting the predictions from each base learner into a 'metalearner', which estimates how much each base learner is weighted in final predictions.[28] Rather than producing final predictions through an average of all base learner predictions, the stacked ensemble estimates a set of weights for the learner. Put in terms of our models, the stacked ensemble can be expressed as:[29]

$$\mathbb{P}(Y = 1|\hat{Y}_{LASSO}, \hat{Y}_{RF}, \hat{Y}_{GBM}) = \alpha_1 \hat{Y}_{LASSO} + \alpha_2 \hat{Y}_{RF} + \alpha_3 \hat{Y}_{GBM} \tag{2.12}$$

where $\alpha_1 \geq 0; \alpha_2 \geq 0; \alpha_3 \geq 0$ and $\sum_{k=1}^{3} \alpha_k = 1$.[30] The strength of the approach, as Van der Laan et al. (2007) and Polley and Van Der Laan (2010) demonstrate, is that the ensemble will perform *at least* as well as the "best" individual model within the ensemble. Across both stages of the stacking process – fitting component models and then estimating weights for those model predictions – all modelling concerns and recommendations about best-practices for cross-validation, loss functions, and parameter tuning apply. Indeed, a popular technique for estimating $\alpha_k$ is through a LASSO, so that regularization lowers the risk of overfitting.

### 2.4.5 Evaluating and Comparing the Stacked Ensembles

After fitting a stacked ensemble for each theory of revision, I compare the final models with precision recall (PR) and receiver operator characteristic (ROC) curves. The area

[26] Technically, the gradient boosting machine and random forest are also a type of ensemble method, because they aggregate predictions across multiple trees. But, the models are different from a stacked ensemble in that they weight all component learners equally.

[27] p605

[28] See Carroll and Kenkel (2016) for a recent application in Political Science.

[29] I use the H2O (The H2O.ai team, 2015) stacked ensemble functionality.

[30] Here I use the notation from Naimi and Balzer (2018), because it is easily interpreted. For a more thorough and formal presentation of the same process see Van der Laan et al. (2007).

under both curves is calculated for *test* sets, with an emphasis on the former because conflict onset is skewed toward zero (Cranmer and Desmarais, 2016; Davis and Goadrich, 2006). The set of variables with the highest test-set AUC score then represents the most empirically-supported theory. Example curves are visualized below in Figure 2.3. Per Davis and Goadrich (2006), the relevant quantities behind the curves are:

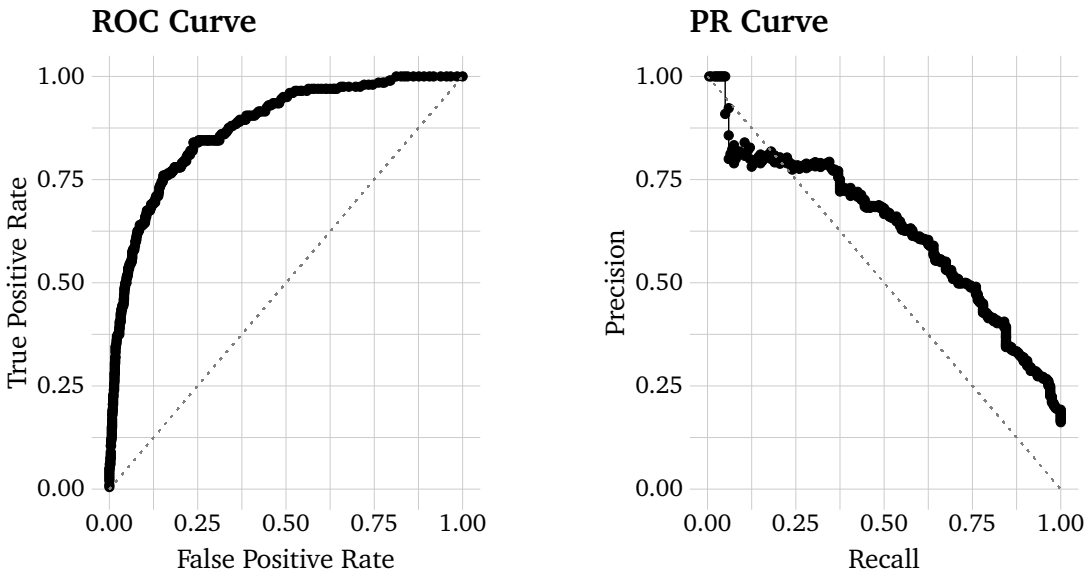
- True Positive Rate: $\frac{TP}{TP+FN}$
- False Positive Rate: $\frac{FP}{FP+TN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$

The PR-AUC and ROC-AUC curves then compare relative model scores for each statistic across different classification thresholds (where a case is classified as 1 at predicted probabilities ranging from 0.0 to 1.0.), which are denoted by the diagonal lines in Figure 2.3. Ultimately, the closer either AUC statistic is to 1, then the better model performance is, but for this data the PR-AUC is a stronger and more difficult benchmark.

Lastly, because of imbalance in the outcome variable – with most country-years lacking any conflict onset – I employ downsampling. Rather than selecting a completely random set of observations to train a model on, where that sample is also characterized by class imbalance, downsampling intentionally selects a lower proportion of the minority class. This creates a training set where the distribution of both classes is even. While this does induce some sampling bias – as the training sample now looks intentionally different from the full dataset – if one has a substantive reason for emphasizing the minority class, then that bias is often worth the increase in predictive accuracy toward the minority class.[31] In our case, a useful model needs to be able to predict cases of onset, because a naive simple model will tend to classify all years as peaceful and be correct most of the time. However, given the importance of preventing conflict, accurate predictions of when conflict does occur are of greater importance.

[31] In the case of these models, while downsampling lowers the test set ROC-AUC it increases the test set PR-AUC, because the latter emphasizes accuracy in the minority class.

Figure 2.3: Precision and Recall Curves, Stacked Ensemble




Example precision and recall curves for stacked ensemble trained on all possible features post-1945. Notably, while the ROC-AUC is 0.875, the PR-AUC is lower at 0.622. The drop in model accuracy is reflective of the class imbalance in conflict outcomes, where a high ROC-AUC is easy to achieve by predicting most cases as never having any conflict. Visually, a model with high predictive accuracy will curve into the top-left corner of the ROC-AUC plot and the top-right corner of the PR-AUC plot.

## 2.5 Results

Before outlining the data and results, we should be careful to note that these ensembles are not meant to be *the one true model of international conflict*. Indeed, a careful read of the results will certainly raise questions of whether the models should have higher accuracy in test sets. But this paper does not intend to capture all of the variance and nuances behind interstate conflict. Rather, the goal is to compare how much revisionist behavior can be predicted, given certain information about a state – whether that be changes in their domestic political environment, differential growth rates, and/or dissatisfaction with the international system. On that point, as Figure 2.4 demonstrates, test set accuracy is the strongest for each time period when every possible variable is included in model training.

Much like the $R^2$ in a linear regression always increasing with more variables, including additional variables adds predictive value. But the figure also makes it clear that not all variables contribute equally, which is our primary interest.

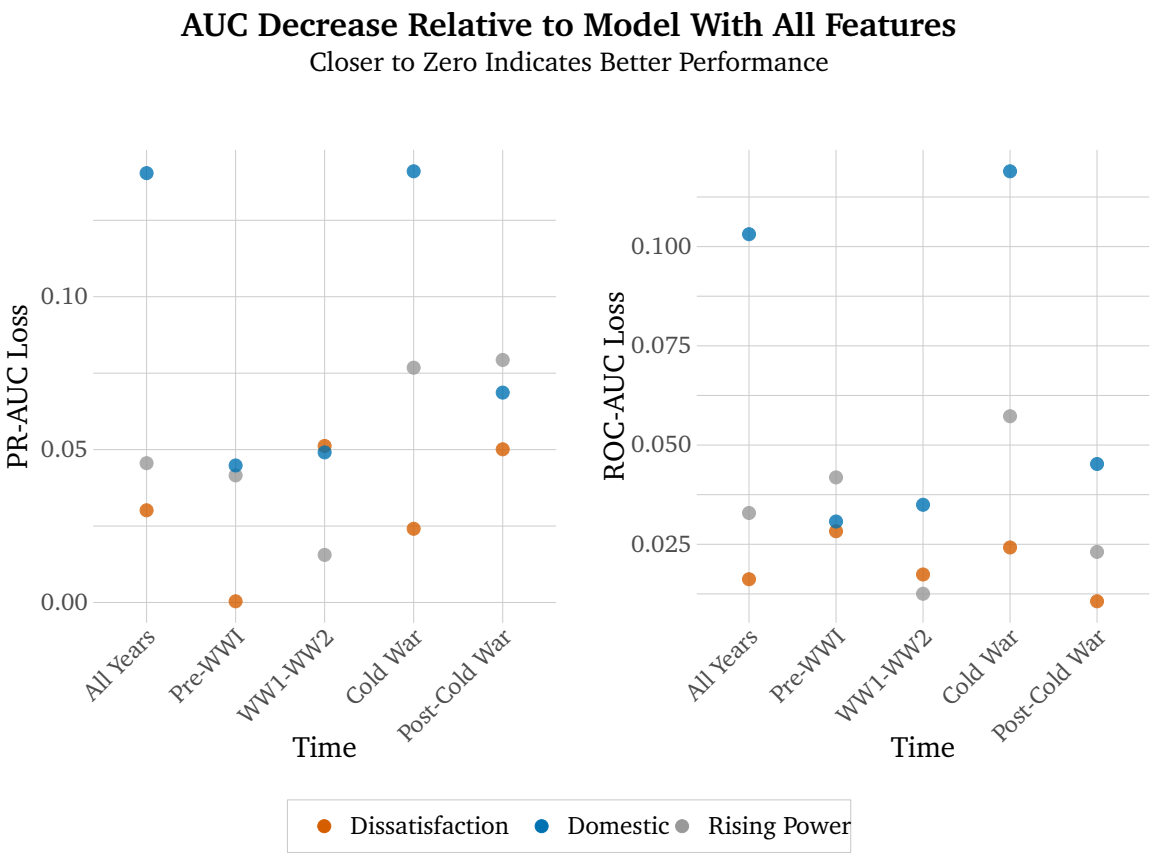
The test-set results for each are displayed below in Figure 2.4. Because the model trained on all possible features always returns the highest predictive accuracy in the test set, I display every other model's predictive *in comparison to the corresponding model trained with all features*. The left-hand plot compares each by their PR-AUC scores while the right-hand plot compares by ROC-AUC scores. For each plot the x-axis represents the time periods considered. The y-axis is then the difference in predictive accuracy for a model trained with variables corresponding to a single comparison, relative to a model trained on that time period's data with all possible features. The closer a model's value is to zero on the y-axis, then the better that model performs because that model is closer to the best-case model, given the available features. Models within each time period are differentiated by their color, with: orange for a model trained on features representing international dissatisfaction, blue for domestic variables, and grey for differential growth rates. Lastly, the first column for both plots includes model results when all years are considered.

The tables in Figure 2.4 show models trained on international dissatisfaction providing generally the highest test set predictive accuracy, with 8 out of the 10 AUC scores having the highest value for models trained on international dissatisfaction.[32] The one exception for both PR-AUC and ROC-AUC scores is between World Wars I and II, where models of rising powers return the highest score. In a sense, this is not surprising, because most interstate conflict during the time periods was initiated by Germany, which for both wars is often referred to as the quintessential rising power. These lessons of both World Wars are understandably applied to many contexts across space and time, though this figure suggests that one should be hesitant to do so. Indeed, dissatisfaction returns a much higher predictive accuracy than rising powers in all other time periods.

One immediate concern when seeing these figures is that when considered separately,

[32]This is not including the models trained on all variables, which unsurprising are the most accurate.

Figure 2.4: Test-Set Accuracy By Time Period and Features




Across all time periods, the stacked ensemble trained with all possible features has the highest predictive accuracy in the test set. This plot compares models trained on a single set of features corresponding to each theory to the model trained with all possible features. The x-axis corresponds to the time period. The y-axis corresponds to how close the model of interest is to the model trained with all features. The left-side figure includes model PR-AUC scores and the right-side figure includes model ROC-AUC scores. Dissatisfaction returns the highest test-set accuracy expect for during the World Wars. This represents the fact that while also dissatisfied, Germany was a quintessential rising power before both conflicts and initiated the majority of the observed conflicts. However, the decrease in relative predictive accuracy for rising powers across all other time periods captures the danger of extrapolating Germany's behavior during the two wars and mapping it onto all other cases and time periods.

dissatisfaction may tend to provide greater predictive accuracy than differential growth rates and domestic political changes, but the most accurate model – including all variables – may rank the variables in an entirely different manner. In order to investigate whether or not models trained on all possible features weighs the features in the same manner as

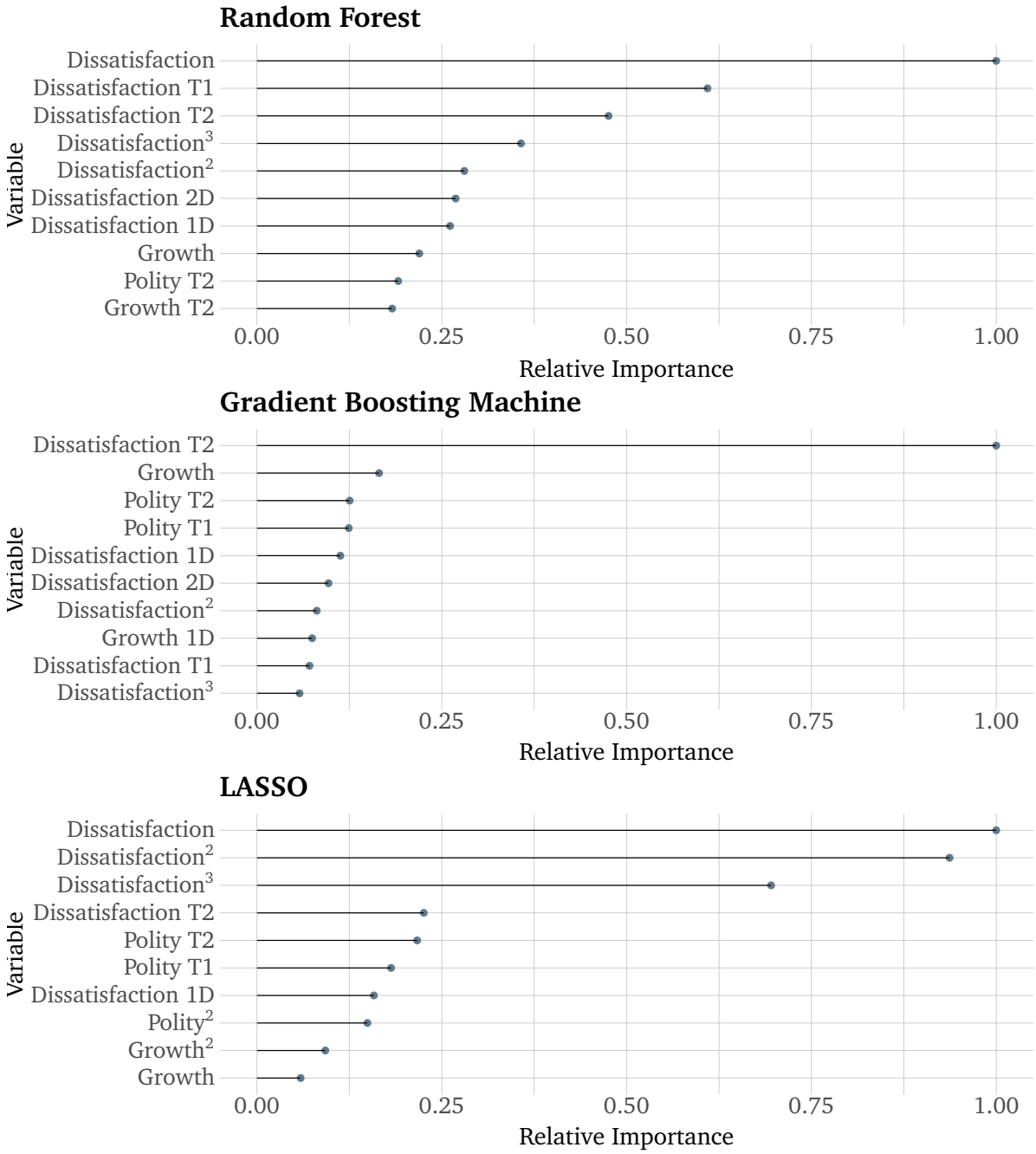
Figure 2.4 would suggest, I include variable importance plots. Fortunately, when we start to open the hood and investigate how the aggregated model processes the data, then we see that each of the machine learning models within the stacked ensemble gives dissatisfaction primary importance. Figure 2.5 plots out the top-10 variables in terms of importance for each model when trained on all possible years.[33] Variable importance plots capture a variable's prominence within a model and how much a variable's inclusion tends to reduce the model's mean squared error (or other loss function) on training data. The x-axis in Figure 2.5 is a variable's importance scaled between 0 and 1 and the variables are ranked along the y-axis in descending order from most important and down.

The variable importance plots all emphasize dissatisfaction's prominence in model-training. For the random forest, dissatisfaction and its transformations are the most important variables. For the gradient boosting machine, dissatisfaction lagged by two years has the most importance (by quite a bit) and for the LASSO dissatisfaction is similarly prominent across the top variables. Coupled with models trained on dissatisfaction returning the generally highest AUC scores in test sets, dissatisfaction's importance in models trained on all features gives it strong empirical support. However, neither AUC curves nor variable importance plots tell us anything about the estimated *directionality* of the relationship between dissatisfaction and conflict onset and that relationship's relative linearity. Fortunately, one straightforward way to capture this is to produce predicted probabilities across a reasonable range of values for all other features and to see how the predicted probability of conflict onset varies across possible values for dissatisfaction.

For complex non-linear machine learning models, partial dependence plots are a useful solution to challenges of interpretability (Greenwell, 2017). A partial dependence plot (PDP) holds all observations at their observed value and records the predicted value for each observation, varying one feature across a predefined range. Here, I take the model fit with all features in the Post-Cold War era, holding all features at their observed value except for dissatisfaction in the current year. Then for each observation the predicted probability of

[33]The rankings are consistent when mdoels are trained on the different time periods.

Figure 2.5: Variable Importance Plots




Top ten features per model, *in training* when each model is trained on all possible features for all possible years. Relative importance presents a standardized measure of how much a model's loss-function tends to decrease when a variable is included into a model.

conflict is recorded at each value of dissatisfaction.[34] While the complexity of these machine

[34]I define the possible range of values for dissatisfaction as the range between the minimum and maximum observed measure of dissatisfaction for any country in the years under consideration.

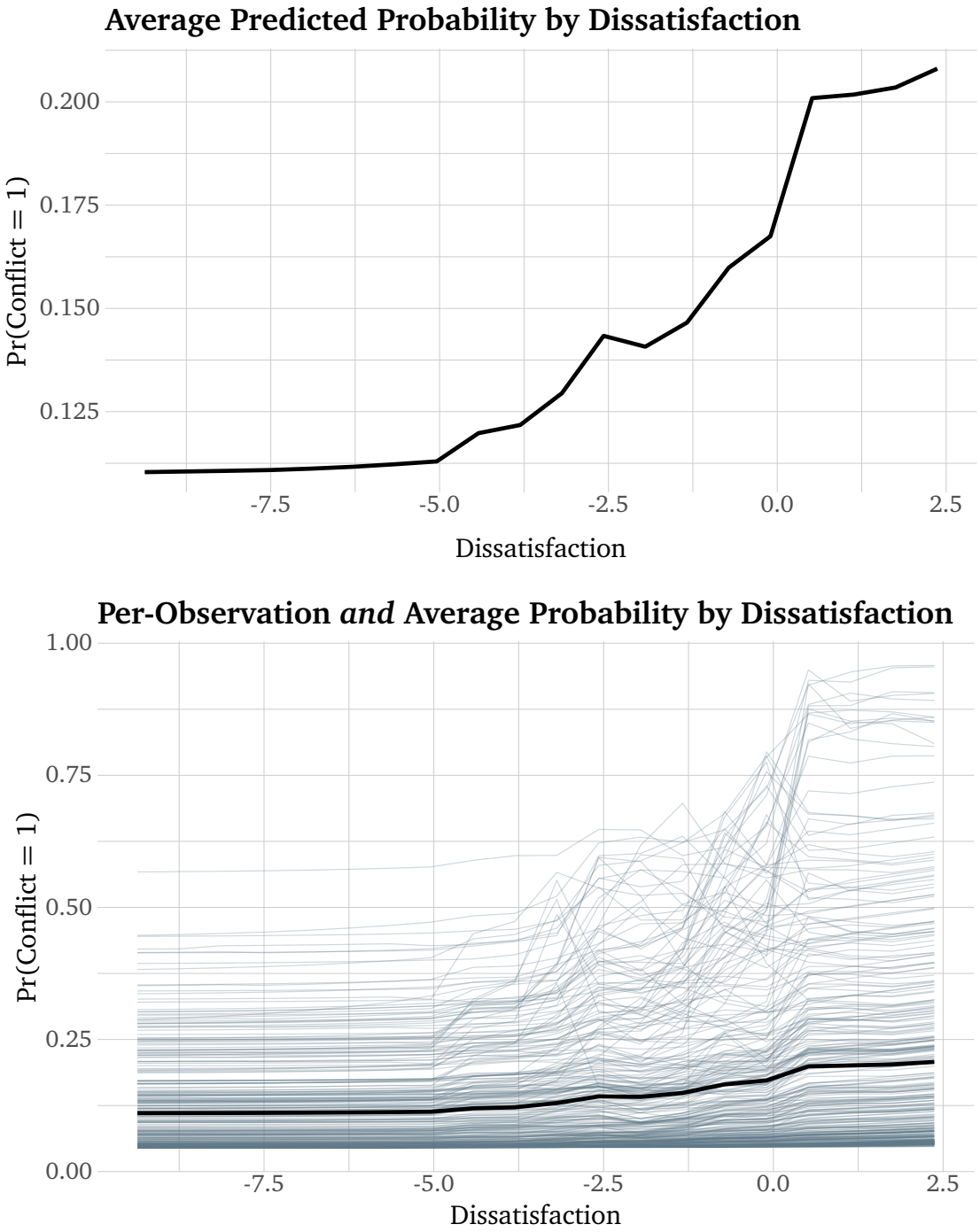
learning models is a barrier to comprehensively mapping predicted probabilities across all possible situations, partial dependence plots do give us a fairly representative interpretation of how the model of interest is actually employing dissatisfaction in the actual data at hand, since predictions for each observation are included.[35]

Figure 2.6 provides two PDPs. The top figure is solely the average predicted probability of conflict onset for all observations, at each possible value of dissatisfaction. As we can see, while non-linear, the relationship is positive, confirming that the stacked ensemble does indeed tend to associate dissatisfaction positively with the probability of conflict onset. Next, visualizing how this average relationship is decided upon, the bottom plot provides a fairly comprehensive summary of the data-generating process for interstate conflict. The average association is positive, but we also see the model capturing a substantial amount of variation across observations, with a highly-variant intercept term and slopes. While most observations are deemed low-risk of conflict onset regardless of their dissatisfaction – capturing the class imbalance in conflict onset – we see the proportion of cases that are considered high-risk starting near a predicted probability of conflict at 0.5, even for the minimum possible value of dissatisfaction. These high risk cases also have a more severe slope, quickly ramping up to a probability of conflict onset near 1 as dissatisfaction increases – which we can see in the plot's top-right corner. Understood in terms of interstate conflict, most states in most years do not consider starting a military conflict as a possibility, but for a small handful of states the prospects are very real and quickly can escalate.

Lastly, unpacking the machine learning model's complexity, I also consider how the model treats observations which are both highly dissatisfied and a rising power. In Figure 2.7, I calculate the average predicted probability of conflict onset for all observations, varying both dissatisfaction and differential growth rates. While the probability of conflict onset clearly varies the most around dissatisfaction, with the average predicted probability not varying by differential growth rates unless dissatisfaction is sufficiently high, we do

[35]The danger of these plots is that they are easily discussed in a causal manner, but for the models at hand we are maximizing predictive accuracy, which is often not the same as a research design for reaching causal identification. Nonetheless, we do receive a measure of model-based association for any variable of interest.

Figure 2.6: Predicted Probability of Conflict Onset By Dissatisfaction




Predicted probability of conflict onset by dissatisfaction. The top plot includes the average predicted probability at each value of dissatisfaction. The bottom plot includes the predicted probabilities of conflict onset *for each observation* at each level of dissatisfaction.

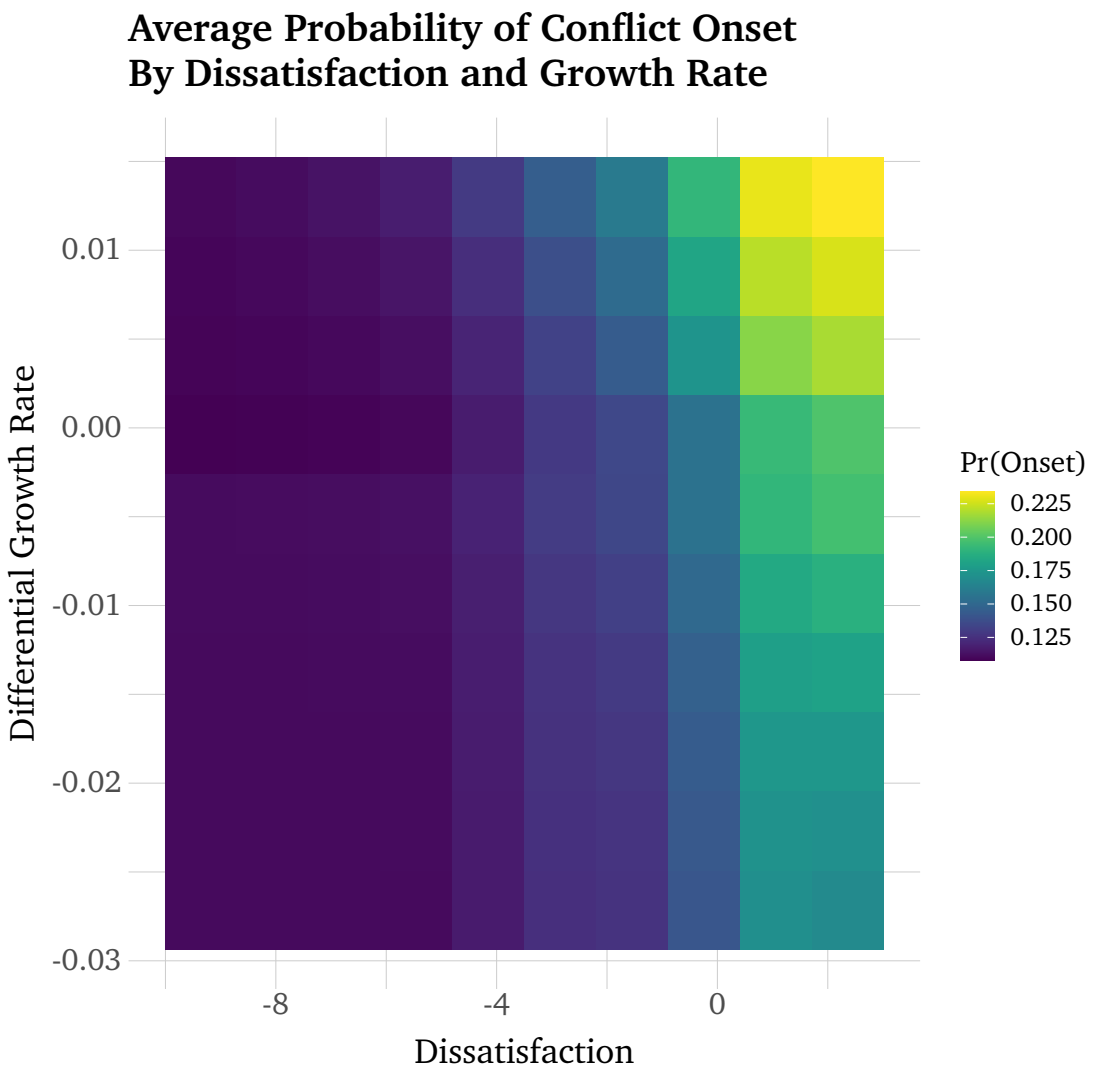
see that the highest-risk cases on average (those in the top-right corner) are states which are both intensely dissatisfied and rising quickly. However, even if a state is not growing or shrinking, the average predicted probability of conflict is relatively high when a state is sufficiently dissatisfied. This figure captures the important point that dissatisfaction is not the *only* predictor of conflict, it just tends to provide greater predictive capacity than the other variables under consideration.

## 2.6 Implications

What are the policy implications of these findings? The predictive nature of these results provides particularly useful insights into current and future trends for great power competition, revision, and conflict. The dissatisfaction measure reveals European powers that are unsurprisingly satisfied with a system that has been immensely beneficial to them since the end of the Cold War. These states are therefore unlikely to become conflictual revisionists. Yet, the remaining non-European global great powers have consistently been dissatisfied with their benefits from the international system, given their power-based expectations. Indeed, even the United States is estimated to have reason to think that the international system could be better designed to reflect its preferences and therefore is likely to continue to engage in some conflictual revisionist behavior, despite its hegemony. Moreover, for United States foreign policy, as debates move from counter-terrorism to great power competition, then we can see that China and Iran have been especially dissatisfied, relative to their peers, informing predictions that they are both similarly likely to pursue conflictual revision and be continual hotspots over their peers.

Given that these estimates stop in 2012, we now know that both China and Iran have been frequently active in military affairs throughout their regions and their revisionist tendencies are now at the forefront of policy discussions and debates. Yet, while dissatisfaction estimates cannot be produced for more recent years – due to the component datasets not being available past 2012 – these trends are likely to be more troubling in 2020 for both

Figure 2.7: Average Predicted Probability of Conflict

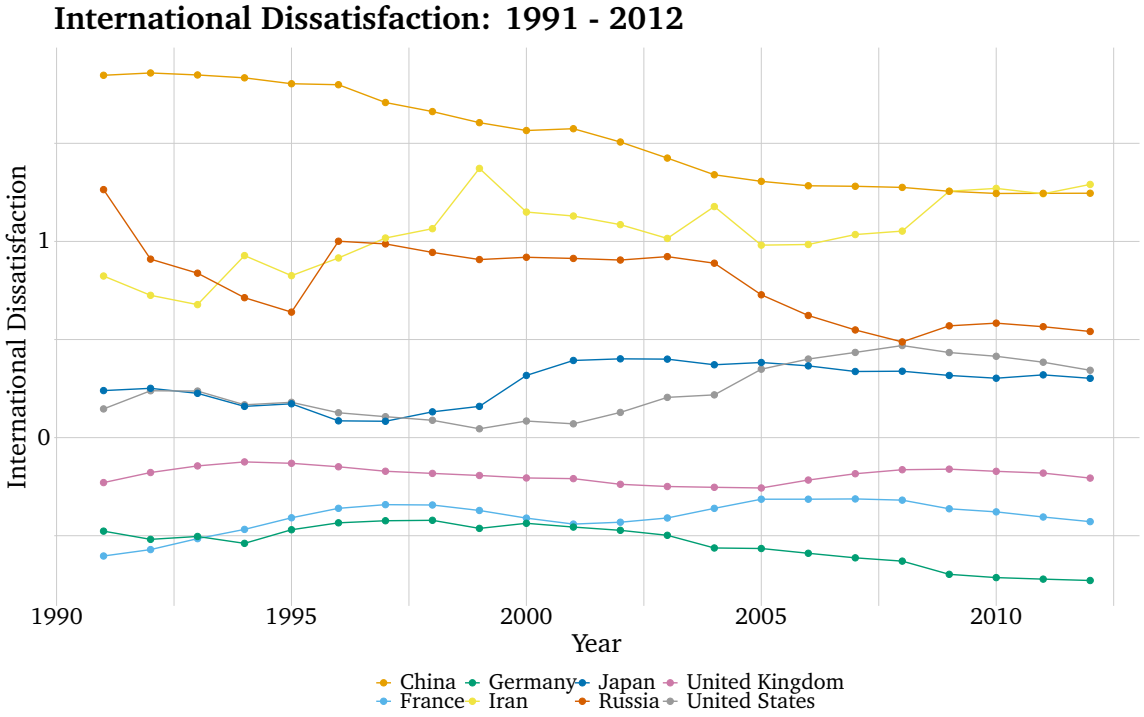



Average predicted probability of conflict onset for all states if each observation is held at its actual values except for the denoted dissatisfaction and growth rate. International dissatisfaction is included on the x-axis and differential growth rates on the y-axis. The lighter colored a grid square is, then the *greater* the predicted probability of conflict onset is. We see the top-right corner with highly dissatisfied and quickly growing states being the highest-risk scenario.

countries. Iran's nuclear program has advanced – further alienating the country and producing increasingly stifling sanctions – and China is increasingly seen as an adversary, not a partner. The estimated measure and predictive exercise therefore highlight China and Iran as being the highest-risk of revisionist behavior. On the other hand, while Russia is

also more dissatisfied than other large states and therefore likely to subsequently initiate conflict – which we retrospectively saw in Ukraine – the measure suggests that China and Iran merit greater focus and expectations for subsequent revisionist behavior than Russia.

Figure 2.8: International Dissatisfaction by Great Powers, 1991-2012




International dissatisfaction estimates for select great powers from 1991 to 2012. Values greater than zero represent increasing dissatisfaction and values below zero represent increasing satisfaction with the international status quo. While the European powers are consistently satisfied with the status quo, non-European states display a consistent desire of and expectation for greater benefits. Iran's international dissatisfaction is particularly striking, likely reflecting the growth of its nuclear program and the intense international sanctioning that has followed it.

Summarizing and discussing the dissatisfaction estimates in more detail, figure 2.8 includes the dissatisfaction estimates for a handful of great powers from 1991-2012. With the European powers below zero, they are estimated to be satisfied with the current state of affairs for each year. However, when it comes to the last year of data – 2012 – we see dissatisfaction from China, Russia, and the United States. Even more troubling – and likely

a foreshadowing of the next 7 years – Iran is the most dissatisfied state by 2012. Indeed, in 2011 the United States congress passed legislation to begin sanctioning foreign banks processing transactions with Iran's Central Bank.[36] Put in this context, as Iran's nuclear program has developed, so too have its relations with many countries frayed. While Iran's estimated dissatisfaction provides an assurance of the measure's validity, it also emphasizes the security dilemma associated with developing a nuclear program. As Iran's program has advanced, so too have other countries grown more wary of Iran's position in the world, creating a greater sense of threat to Iran and perceived need for the nuclear program. Given the state of relations with Iran and intense sanctioning under the current U.S. administration, we should expect to see continued military conflict in the Persian Gulf, not just as a part of the ongoing catastrophic proxy war between Iran and Saudi Arabia in Yemen.

Turning to China, if its dissatisfaction is considered alongside this paper's results, then greater focus should be directed to its relations with other states – or the lack thereof. While the dominant Chinese narrative is likely around rising powers and the 'Thucydides Trap' (Allison, 2017), whether or not China pursues a long-term strategy of revision (Johnston, 2003) may very well be decided more by whether or not other states are comfortable with deepening relations with China as it grows, not whether or not a power transition will occur or if Xi Jinping's leadership will remain uncontested. This is not to say that the latter two questions are unimportant, but this paper's results call for a focus on the rest of the international system and whether that system is likely to behave in a way that China would want to revise. In response to these concerns one may ask why China would seek to revise an international system which enabled its rise. However, a straightforward reply is that China has good reason to think it can and should be getting more from its international environment, based on its power position alone.

For example, as the Belt and Road Initiative (OBOR) expands and East Asian institutions like the Asian Infrastructre Investment Bank (AIIB) are spearheaded by China, should potential partners demonstrate a willingness to buy into those institutions, then large revi-

[36] https://www.armscontrol.org/factsheets/Timeline-of-Nuclear-Diplomacy-With-Iran

sionist conflict will likely decrease in probability. On the other hand, if potential partners view these institutions as illegitimate or too high-risk for the expected payoffs, then China's expected international benefits will continue to outstrip its actual international benefits and radical change to the international system will gain appeal. The appeal of these institutions to potential members is an open, but potentially the most important, question when it comes to China's future foreign policy.

## 2.7 Conclusion

When do revisionist states become revisionists? This paper approaches the question of revision's origins through a predictive approach, comparing the predictive accuracy of machine learning models trained on different theoretically-informed variables when forecasting interstate conflict initiation. While models trained on differential growth rates and domestic political changes demonstrate meaningful predictive accuracy – with the former being the most accurate model during the World Wars, a state's international dissatisfaction presents the most predictive accuracy, in most years. These results do not invalidate theories of domestic politics and rising powers when considering international revision, but they do support concentrating one's focus on a state's relative standing and relations with other states when asking when international revision occurs. In other words, revision does not just follow from changes within the state, but in response to whether or not a state has reason to believe that the system can plausibly be altered to better suit that state's desires.

I also investigate the implications for current great powers relations. Both Iran and China are estimated to be the most dissatisfied major states during the last year of available data. While the two states' dissatisfaction, (alongside its predictive capacity) lend credence for the measure itself, these two states' strong dissatisfaction with the international system raises questions of foreign policy design and potential future outcomes. For both states, dissatisfaction stems from a lack of international standing and strong relations commensurate to their capabilities. This dissatisfaction over their international *relations* with other states

means whether or not the two turn toward increased or decreased international conflict in the future will likely be very predictable by the decisions of other states and whether potential partners are willing to deepen relations with both states. In this context, questions of international revision and conflictual actors should be just as, if not more, focused on the makeup of the relations that constitute the international system as the potential belligerent actors themselves.

Lastly, this project has been relatively agnostic on the sources of international dissatisfaction. A logical next step for future research is to examine why some states become highly dissatisfied and socially excluded whereas others find themselves in a situation of general satisfaction and benefit with the status quo. While domestic politics and growth rates are certainly part of the picture – raising questions of a potentially spurious relationship – as we saw earlier in Figure 2.1, the correlation between dissatisfaction and growth rates is below 0.5 and the correlation with polity scores is almost zero. Put differently, the question is more than one of academic interest and nuance. While this study identifies dissatisfaction as an especially meaningful predictor of interstate conflict and revisionist states, it provides little insight on the best policy levers that can be pulled in order to peacefully bring states into the international fold – remedying intense dissatisfaction among great powers. Understanding why some international environments and states may be more prone than others to social exclusion may then provide a better understanding of which levers are best-suited for ameliorating these complicated but consequential dynamics.

# Chapter 3: Estimating Heterogeneous Spillover Effects

## 3.1 Introduction

In many social settings, researchers are interested in estimating the causal effect of a treatment. These causal estimates are generally produced through an average treatment effect that compares the average outcomes for units that receive treatment and units that do not. While many concerns arise in observational settings when comparing treatment and control groups, two prominent threats to the utility of a straightforward comparison of means are: heterogeneous treatment effects and treatments spilling over from one unit to another. In cases of the former, given sufficiently heterogeneous effects, estimates of average effects will not map onto the actual group-level effects. In cases of the latter, comparing treated to untreated units will return a biased estimate, lowering an effect estimate because observations in the control group will be mistakenly labelled as having no treatment exposure of any kind.

While techniques have been developed for both situations separately, what if spillover effects are present *and* they are heterogeneous? In the context of international politics, Jervis (1998) argues that in complex systems such as the international, where units are heavily interconnected, "we can never do merely on thing." Put differently, in social settings, pulling one lever is likely to have unforeseen additional effects in contexts not initially considered, where those effects will also vary in their magnitude. Across applications such as get-out-

the-vote (GOTV) programs, foreign aid provision, and school programs, spillover effects are a widely-recognized reality, where observations tied to the treatment group in some way receive indirect treatment exposure. In addition, these applications often exhibit heterogeneous treatment effects, with the treatment's magnitude and direction varying widely across contexts. It therefore follows that, across domains, spillover effects should not only be considered, but also checked for effect heterogeneity. Yet, while heterogeneous treatment and spillover effects are methodologically important and potentially simultaneously present in many contexts, little guidance exists on how the two can be combined in a causal inference framework.

This paper discusses and presents an approach that combines recent advances in both literatures. The proposed method works in three steps: defining potential outcomes in terms of both direct and indirect treatment exposure, weighting observations by their probability of receiving indirect treatment exposure, and then incorporating these potential outcomes and sample weights into a causal random forest. In terms of disaggregating potential outcomes, Aronow et al. (2020) demonstrate that once units are differentiated by their treatment status and whether they share a social tie to a treated unit, then indirectly exposed units can be compared to units with no indirect exposure, providing a causal estimate of the sample's spillover effects. This comparison is equivalent to calculations of an average treatment effect if no spillovers are present, meaning comparing groups according to their spillover conditions is compatible with methods for identifying heterogeneous treatment effects. (Imai and Ratkovic, 2013; Wager and Athey, 2018, e.g.,) The two primary differences between the proposed and existing heterogeneous treatment effect estimation procedures are that 1) indirect treatment exposure is substituted for directly receiving treatment and 2) the two groups under comparison are control group units in an experimental setting that are indirectly exposed and unexposed.

After reviewing the relevant methods separately, I further discuss how the two are compatible theoretically and in practice. After formally discussing the relevant measures, I demonstrate how available R packages can be used in tandem to estimate conditional av-

erage spillover effects, coupled with a monte carlo simulation to demonstrate the software combination provides accurate results. The simulation also returns strong support for including sample weights for the probability of indirect treatment exposure when estimating a causal forest for spillover effects. Lastly, I apply the combined methods to a study on spillovers from anti-bullying programs in school settings, where treatment effects are heterogeneous across schools, demonstrating how a pilot study can inform expectations of where subsequent efforts will be effective and where they are likely to produce undesired outcomes. The application also demonstrates how outcomes can be disaggregated to differentiate between trends due to direct and indirect treatment.

## 3.2 Related Work

This paper seeks to synthesize, from a causal inference perspective, two popular methodological research topics: heterogenous treatment effects and spillover effects. In many contexts, spillover effects and effect heterogeneity are plausibly both present, making accurate estimation important for academic research, product design, and policy implementation alike. Both methodological literatures are fast-growing and tend to approach estimation procedures from the potential outcomes framework, where causal effects are understood to represent the difference between a unit's outcome if that unit does and does not receive treatment. More specifically, the two recent methodological developments – with accompanying software – that inform this paper's proposed approach are Aronow et al. (2020) and Wager and Athey (2018). From a technical perspective, this article's contribution is that it provides a theoretical and applied guide to combining the methods effectively. Indeed, while developed separately, the two methods can be viewed as complementary. I formally discuss the two approaches in turn before discussing how the two methods can be synthesized and used together.

### 3.2.1 Spillover Effects

A fundamental concern when estimating causal effects is the presence of "interference" between units, a class of instances where one unit's outcome is not only a result of its treatment status, but also the treatment status of other units. (Hudgens and Halloran, 2008; Rubin, 1990; Taylor and Eckles, 2018; VanderWeele and Tchetgen, 2011) One way that interference occurs, and this paper's focus, is through spillover effects, where indirect treatment exposure is conferred through a tie linking two units.[1] In these situations, treatment literally spills over from one unit to another through their shared tie. These spillovers can be geographic[2], social[3], or through linked processes in a complex system.[4] I mention these different settings because, while I discuss spillover effects specifically in terms of social networks throughout the rest of this paper, the proposed framework is readily adjustable to other contexts.

This paper's proposed method draws primarily upon Aronow et al. (2020), who frame spillovers through four experimental conditions. Each unit's condition can be derived as long as treatment assignment and the adjacency matrix of ties between units are known. Given knowledge of ties and treatment, units can be differentiated based on whether they received treatment and/or indirect exposure through a potential spillover. Aronow et al. label these conditions the "exposure mapping", with four possible conditions:

- $d_{11}$: Direct and indirect treatment exposure
- $d_{10}$: Isolated direct exposure
- $d_{01}$: Indirect exposure
- $d_{00}$: No exposure

Under these conditions, indirect treatment exposure occurs when a unit shares a networked tie with a unit that received direct treatment exposure. A unit's potential outcome is then

[1] Recent examples of studies estimating spillovers include Baicker (2005); Cheung and Ping (2004); Haushofer and Shapiro (2018); Jones et al. (2017); Ng (2000); Nickerson (2008); Paluck et al. (2016)

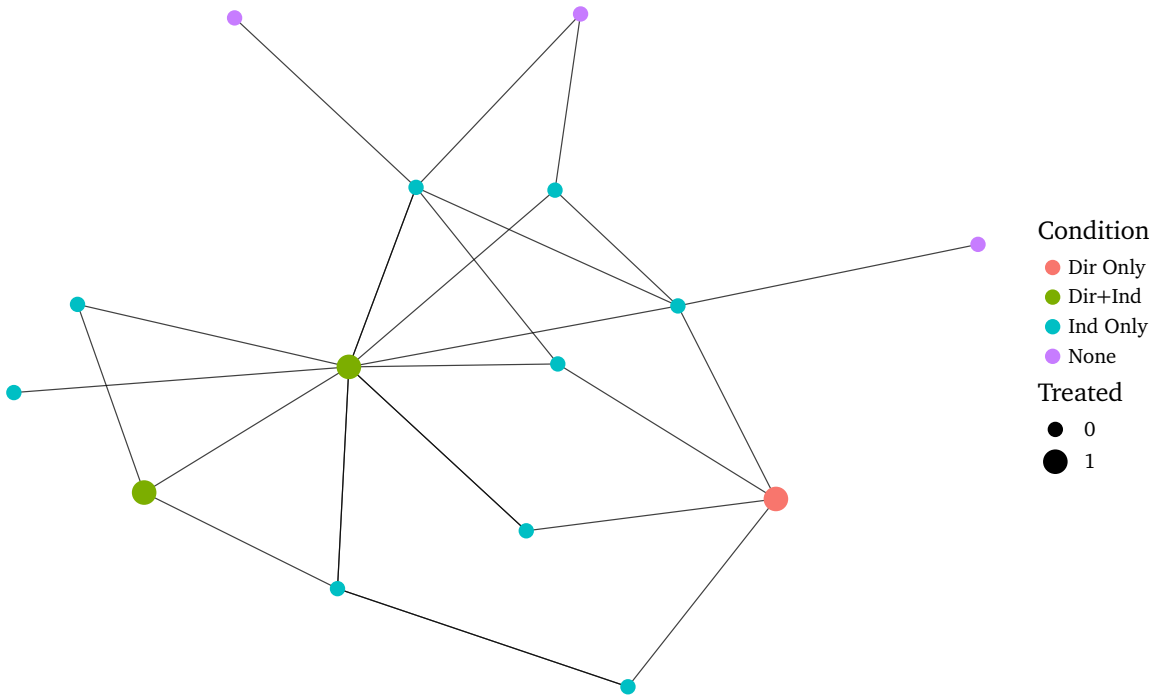
[2] E.g., a change in one location impact neighboring locales.

[3] A change to one person can impact their friends and family.

[4] Consider the recent public health pandemic, which was shortly followed by a massive drop in oil prices.

labelled $y_i(d_k)$, as opposed to the common notation $y_i(1)$ and $y_i(0)$ for whether a unit is in the treatment or control group. For example, in the later example of spillover effects from anti-bullying programs in schools, indirect exposure occurs when a student is a friend with a student that was assigned to an anti-bullying program, where the anti-bullying program is the treatment of interest. However, in order to estimate these indirect effects, both the school's friendship network and the students assigned to the anti-bullying program must be known.[5]

Figure 3.1: Example Network: Treatment Status and Exposure Condition




Example network, where larger nodes are the treated units and ties are pathways through which indirect exposure can occur. Node color represents the exposure condition, spanning: no exposure (purple), indirect exposure only (light blue), direct treatment only (red), and both direct treatment and indirect exposure (green).

The categories then allow the estimation of $\tau(d_k, d_l)$, or the causal effect of being in exposure condition $d_k$ rather than exposure condition $d_l$. In terms of spillover effects, two

[5]Or the social network and treatment assignments must be estimable from the available data.

comparisons are especially relevant: 1) $\tau(d_{01}, d_{00})$ – comparing units with indirect and no exposure in the control group and 2) $\tau(d_{11}, d_{10})$ – comparing units in the treatment group that also receive indirect exposure to those only with direct treatment. These comparisons allow us to move from only comparing average outcomes in the control and treatment groups to also comparing units within each direct treatment condition based upon indirect exposure. The most straightforward version of these spillover effects, and my subsequent focus, is on the first comparison: units in the control group with indirect exposure and units in the control group with no exposure. However, for any comparison of groups $k$ and $l$, then Aronow et al. demonstrate that the average causal effect can be estimated with:

$$\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l) = \mu(d_k) - \mu(d_l). \tag{3.1}$$

where $\mu(d_k) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k)$ is the average potential outcome for units with exposure $k$. Equation 1 is therefore a difference of means, just with a more fine-grained set of potential outcomes than the usual formulation of an average treatment effect.

In most settings, however, even if treatment is assigned at random, the preexisting social ties are not random. This means the probability of $y_i(d_k)$ is not equal across all units. As a solution, Aronow et al. (2017) propose estimating $\boldsymbol{\pi}_i$, or observation $i$'s *generalized probability of exposure*. For our framework, $\boldsymbol{\pi}_i = (\pi_i(d_{11}), \pi_i(d_{10}), \pi_i(d_{01}), \pi_i(d_{00}))$. Once these probabilities are estimated, they can then be used as weights for the observed outcomes, approximating "as if" random assignment to the observed potential outcome for each unit.[6] In order to estimate $\boldsymbol{\pi}_i$ for each unit, Aronow et al. (2020) provide software that calculates the range of possible treatment assignments, denoted $\Omega$, and then calculates the proportion of times that each unit $i$ finds itself in exposure condition $d_k$ for the observed network across the set of possible treatment assignments.[7] Given a sufficiently large set of possible treatment assignments, where $|\Omega|$ is high, treatment vectors are sampled finitely

[6] This approach is similar to that employed in Ugander et al. (2013).

[7] See https://github.com/szonszein/interference for the exact code and an example on p11 of Aronow et al. (2020).

to approximate the range of values in $\Omega$.[8]

The resulting probability estimates of each unit $i$ being in exposure condition $d_k$ can then be used for the Horvitz-Thompson inverse probability estimator:

$$\widehat{y_{HT}^{T}}(d_k) = \sum_{i=1}^{N} \boldsymbol{I}(D_i = d_k) \frac{Y_i}{\pi_i(d_k)}. \tag{3.2}$$

Once each unit's observed outcome is inversely weighted by its estimated probability of being in the observed exposure condition, we can then combine Equations 3.1 and 3.2 to produce:

$$\widehat{\tau_{HT}}(d_k, d_l) = \widehat{\mu_{HT}}(d_k) - \widehat{\mu_{HT}}(d_l) = \frac{1}{N}\left[\widehat{y_{HT}^{T}}(d_k) - \widehat{y_{HT}^{T}}(d_l)\right]. \tag{3.3}$$

where $\widehat{\tau_{HT}}(d_k, d_l)$ is the Horvitz-Thompson weighted estimate for the average causal effect of being exposed to $k$ rather than $l$. In this sense, Aronow et al. (2020) do not only argue that spillover effects can be estimated through the familiar difference-of-means framework once exposure conditions are calculated, but that the observed potential outcomes should also be inversely weighted by probability estimates that each observation receives its observed treatment condition.

Because Equation 3 is a difference of means with added weights, it is readily applicable to the methods I discuss next about heterogeneous treatment effects. Across techniques, heterogeneous treatment effect estimators calculate the difference of means across potential outcomes, conditional on covariate values. However, I also find in monte carlo simulations that applying the Horovitz-Thompson weights from Aronow et al. (2020) almost always increases the accuracy of heterogeneous treatment effect estimators when they are applied to spillover effects. I discuss the proposed method for heterogeneous treatment effects next.

[8]While these probabilities are currently estimated only using the observed ties, an interesting next step in this literature could be estimating exposure weights by directly modelling a network's tie-formation using tools such as exponential random graph models (ERGMs) or other inferential network analysis tools. (e.g., Cranmer et al., 2017; Hunter et al., 2008; Minhas et al., 2019; Victor et al., 2017; Wilson et al., 2017)

### 3.2.2 Heterogenous Treatment Effects

Machine learning methods are increasingly popular tools for uncovering the presence of heterogeneous treatment effects, with a recent focus on forest-based algorithms. (Athey and Imbens, 2016; Green and Kern, 2012; Grimmer et al., 2017; Hill, 2011; Hill and Su, 2013; Hill et al., 2020; Wager and Athey, 2018) Generally speaking, there are two approaches to modeling effect heterogeneity. Given a theoretically informed set of moderators (James and Brett, 1984), one can denote and compare outcomes across subsamples or fit a regression with an interaction term. (Braumoeller, 2004; Brambor et al., 2006; Esarey and Sumner, 2018, e.g.,) Alternatively, a situation of interest may be characterized by unknown heterogeneity across an unknown number of variables. Here, machine learning algorithms can be used to sort through the possible moderators with either the aforementioned forest-based methods or a form of variable selection (Imai and Ratkovic, 2013), using the available data to assess moderator plausibility. Considered this way, machine learning approaches to uncovering effect heterogeneity can be incredibly valuable for taking a pilot study or A/B test and uncovering where the proposed policy or product change will likely have the largest or smallest effects when implemented broadly in one's population of interest.

This paper builds off Athey and Imbens (2016) and Wager and Athey (2018), who develop a procedure for estimating heterogeneous treatment effects with "causal forests" fit through a two-stage 'honest' approach. In the honest approach, when training a model, half of the data is used to construct each tree's splits and the other half of the data is used to assess the causal effects assigned to observations that fall in each tree's terminal leaves. While this comes at the cost of not using all available information when building the structure of the component trees, Wager and Athey demonstrate two substantial benefits in return. First, the honest approach enables observation-level standard errors, which is new development in the random forest literature. Second, it lowers the probability of overfitting, which in this framework occurs when noise-driven outlier observations have disproportionate influence on the final predicted causal effects. Importantly, on the latter,

random forests are generally recognized to have a tendency to mistake high-variance noise for the data's true systematic trend because doing so decreases training set predictive error. In response, calculating each leaf's predictions on data not used for training lowers this risk, because outlier observations can only influence one portion of the training process, not both. For example, while an outlier observation may disproportionally influence the covariate values where splits occur, it will not subsequently inform the prediction made for out-of-sample observations that fall under the resulting leaves. Indeed, this 'honest' aspect of the model is helpful beyond causal forests and has been demonstrated to improve test-set predictive accuracy for a range of forest-based algorithms.

Beyond the honest sampling approach, causal forests are different from random forests because they are not used to maximize within-leaf predictive accuracy. Rather, splits are chosen on covariate values that maximize the difference between the average treatment effect estimates in the two resulting nodes. As a short aside, in a random forest, a model's accuracy can be evaluated using test set data where true outcomes are known and can be compared to predictions, allowing a straightforward verification of a model's predictive accuracy. However, causal effects are never observable because of the fundamental problem of causal inference, so the causal forest's predictive accuracy cannot be confirmed by comparing unit-level predicted causal effects to the true causal effects, which are unobserved.[9] Returning to the causal forest's mechanics, by differentiating average treatment effects across nodes through covariate splits, the causal forest estimates conditional average treatment effects (CATEs):

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right] \tag{3.4}$$

giving us a difference of conditional means at a proposed covariate value.

Formally, the method begins with n units, where a tuple $(X_i, Y_i, W_i)$ is observed with: feature vector $X_i \in \mathbb{R}^p$, response $Y_i \in \mathbb{R}$, and treatment assignment $W_i \in \{0, 1\}$. The causal

[9]This discussion is meant to raise a sense of empirical caution with these models. While they are verified to work in simulation, their predictions should be carefully assessed.

forest then estimates heterogeneous treatment effects by fitting a series of decision trees under the following steps:[10]

1. Draw a random subsample with replacement from the dataset
2. Split the root node into child nodes repeatedly, with the following steps:
   - Select a random subset of variables for splitting
   - At each variable $x$, possible values for splitting, $v$ are considered, with each potential split $(x, v)$ evaluated by how much it increases heterogeneity in the resulting CATE estimates across leafs.
   - Observations within the splitting variable $x$ less than or equal to $v$ are placed in the left child node and values greater than or equal to $v$ in the right child node.
   - If a node lacks valid splits or splits do not improve fit, a node is not split any more and is considered a leaf.

However, since we are working in the honest framework, the independent half of the training dataset not used for model building then determines the predicted outcomes in the terminal nodes for each tree, where the predicted CATE for any observation in a node is the average difference between its control and treated units. These leaves allow for estimating complex moderating effects for each observation, producing a final set of unit-level predicted causal effects, determined by where that observations falls in the causal forest. If one is more interested in how effects vary across groups, then these observation-level predictions can be aggregated by groups and then compared.

More formally, when building the forest, Wager and Athey denote the conditional average treatment effect of a potential split at $X = x$ as:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}}^{Y_i} - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}}^{Y_i} \tag{3.5}$$

[10]This explanation draws on the accessible tutorial at https://grf-labs.github.io/grf/REFERENCE.html.

where $Y_i$ is the outcome, $W_i$ is treatment assignment, and $i \in L(x)$ are indices for the data $X$ which decide the split. $|\{i : W_i = 1, X_i \in L\}|$ and $|\{i : W_i = 0, X_i \in L\}|$ denote the number of observations in both conditions (treated vs untreated for the proposed covariate split). Once the ensemble of trees are fit, the causal forest takes the aggregated ensemble, where each tree's estimated causal effect for a covariate condition, $x$, can be expressed as $\hat{\tau}_b(x)$ and averaged for the whole causal forest, giving us:

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x). \tag{3.6}$$

Equation 6 is therefore the estimated causal effect of receiving treatment (or spillover in our context), conditional on covariates $X = x$.

In summary, the causal forest estimates the difference between each observation's potential outcomes, based on its covariate values. This is accomplished by building a random forest where each trees' splits are decided upon by splitting at the covariate value that maximizes the difference between the average treatment effect in each resulting node. Each tree within the final ensemble provides a treatment effect estimate for observations that fall within the final leafs, where the causal forest makes final predictions by averaging each tree's predictions. These final estimates are produced at the unit-level, where each observation's estimated causal effect of treatment assignment is calculated by the final leaf it is placed in within each of the component regression trees. In the next section I outline how this approach can be merged with the aforementioned discussion of spillover effects to estimate heterogeneity in spillover effects.

## 3.3 Method

In this section I discuss how the two previously discussed approaches to spillover effects and heterogeneous treatment effects are directly compatible and propose a workflow for combining the two in applied research. First, I demonstrate how the two are at their core a comparison of average outcomes across two groups, differentiated by a causal variable

(which can be direct treatment or indirect treatment exposure through a networked tie). Second, I discuss the implementation of the Horvitz-Thompson inverse probability weights in a causal forest, accounting for spillover assignment being non-random, even if treatments are assigned at random. Lastly, I discuss how to evaluate output from a causal forest in terms of spillover effects and their heterogeneity.

Spillover effects and heterogeneous treatment effects are both at their core about comparing means across groups that do or do not have a causal variable (treatment or spillover). Indeed, although the notation in the aforementioned papers is more complicated, due to the nuances of the groups being compared and estimating sample weights, the core formulation of an average treatment effect is the foundation of both approaches:

$$\text{ATE} = \mathbb{E}[y|t = \text{treatment}] - \mathbb{E}[y|t = \text{control}]. \tag{3.7}$$

In the potential outcomes framework, an average treatment effect takes the average outcome in the treatment group and compares it to the average outcome in the control group. With both potential outcomes being unobservable simultaneously at the observation-level, comparing these two averages across treatment groups is recognized as providing an unbiased and useful estimate of the true causal effect of a treatment variable, if treatment is assigned at random and effects are relatively homogenous across units. Although situation-specific methodological challenges drive which groups are compared and how those groups are distinguished, the core calculations always take the difference of means across groups with and without a causal variable.

Keeping this comparison of average outcomes in mind, we can see Equation 3.7 in the more complicated Equations 3.1 and 3.5. Restating in turn, Equation 3.1 is the primary formulation for spillover effects:

$$\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l)$$

and Equation 3.5 denotes calculations for heterogenous treatment effects:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}}^{Y_i} - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}}^{Y_i}.$$

For both of these equations, the effect of interest is comparing the average outcome across groups that do and do not receive some causal assignment. In the spillover equation, $d_k$, is the group with indirect treatment exposure and $d_l$ receives no indirect exposure. For the heterogeneous treatment effect equation, $W_i = 1$ denotes receiving treatment and $W_i = 0$ denotes being in the control group. Both then are averaging outcomes, $y_i$, across the two groups.

Considering how the two equations can be combined, let both $d_k$ and $d_l$ be units in the control group, varying by whether they share a network tie with treated units, conferring indirect treatment exposure. Then the average causal effect of spillover exposure is the difference between the average outcome in both groups $k$ and $l$. Incorporating Equation 3.3, we can then restate this difference in means to be the difference in weighted averages, taking each unit's probability of indirect treatment exposure into account. Moreover, in equation 3.5, if $X_i \in L$ is restated as all covariate values and $W_i$ denotes spillover exposure, rather then treatment exposure, ***then the two equations will return identical outcomes***. This is the core argument for this paper: *because the two methods start with the same methodological assumptions and framework, the two can be used in tandem.*

Next, I outline a series of steps with R code, demonstrating how to combine the methods in practice. In order to do so, the following data is necessary: a network of ties between units, treatment assignment, and observation-level covariates to condition on. Fortunately, recent software developments for modeling heterogeneous treatment effects allow for a straightforward implementation of sample weights, allowing the use of the Horvitz-Thompson inverse probability weights.

### 3.3.1 R Procedure

Before turning to simulations and an application, this section outlines the necessary R code for combining the interference and grf packages. Both packages are recent improvements in the estimation of spillover effects and heterogeneous treatment effects, respectively. After installing the libraries, this code assumes the following objects in R:

- `net`: network object
- `treat`: vector with each node's treatment assignment
- `X`: covariate values for each node
- `y`: outcome variable for each node
- `W`: binary variable denoting whether or not a node receives indirect treatment exposure

Once these objects are defined, the following code can be easily applied to one's data with minimal adjustments.

First, we extract the adjacency matrix from our network of interest:

```
# Extract adjacency matrix from network object
adjmat <- treat_schools_net(net)
```

Then we calculate the probability that each node ends up in each of the four exposure conditions outline in Aronow et al..

```
# Store treatment assignment for each node
d <- make_exposure_map_AS(adjmat, treat, hop = 1)
# hop = 1 for single tie conferring indirect exposure

# Create vectors of different treatment assignments
potential_tr_vector <- make_tr_vec_permutation(
    N = nrow(d), # number of observations
    p = p_treat, # exogenously set probability of treatment
    R = 30, # number of treatment vectors to generate
    seed = 123 # random seed for replication
)

# Calculate exposure probabilities
obs_prob_exposure <- make_exposure_prob(
    potential_tr_vector, # treatment assignments
```

```
    adjmat, # network adjacency matrix
    make_exposure_map_AS, # exposure mappings
    list(hop = 1) # spillover with 1 tie
)

# Exposure conditions
exp_probs <- t(make_prob_exposure_cond(obs_prob_exposure))
```

Now that we have the probability of exposure for each condition for each node, we can apply those probabilities as weights in the causal forest. For this causal forest, we will estimate $\tau_i(d_{01}, d_{00})$, or the expected difference between indirect exposure and no exposure for each unit, conditional on its covariate values. This code fits and evaluates the causal forest on all available data, but can be easily amended for the training-test framework if one wishes to approximate the process of using a pilot study to make predictions about expected effects on new data. Here, let `y` be the outcome of interest, `X` is a matrix of covariates, and `W` is the treatment (here a spillover). Notably, the causal forest includes the option to also match by inverse-propensity scores, estimating within-leaf treatment effects by matching units together based on their estimated probability of receiving treatment. In light of King and Nielsen (2019), I eschew this option and only use the inverse-probability estimates as sample weights, rather than matching on them.[11]

```
cf <- causal_forest(
    X = as.matrix(X), # Covariates
    Y = y$outcome, # Outcome
    W = W$spillover, # Spillover assignment
    sample.weights = 1/exp_probs$ind, # Probability weight
    seed = 123 # random seed for replication
)
```

We can then test for effect heterogeneity in the entire causal forest with the following command:

```
test_calibration(cf)
```

If the term `differential.forest.prediction` in the resulting summary table is positive and significant, then that implies support for effect heterogeneity and we can reject

[11] For more information on implementing a causal forest, https://www.markhw.com/blog/causalforestintro is a helpful resource.

the null of homogenous effects. If the null of homogenous effects has been rejected, then we can use a variable importance plot to see which variables tended to produce the most effect heterogeneity when used to generate splits in the causal forest.[12]

```
cf %>%
    variable_importance() %>%
    as.data.frame() %>%
    mutate(variable = colnames(cf$X.orig)) %>%
    arrange(desc(V1))
```

If the estimated spillover effects are heterogeneous and certain variables are highlighted by the variable importance plot, then we can turn to using observation-level effect estimates to visualize the heterogeneity of interest. If no new test set data is entered, then the default format is to use 'out-of-bag' (OOB) predictions, where each observation's effect estimates are made using trees where that observation was not used for training. Because random forests build each tree using a random sample of data, the expectation is that each data point is not used for building every tree, meaning there are certain trees where the data point approximates a new test set observation that the model is unfamiliar with. Notably, a benefit of the honest approach is that the model produces observation-level standard errors, which `estimate.variance = TRUE` allows.[13]

```
# If no test set, using OOB predictions
predict(cf, estimate.variance = TRUE)

# If there is a test set
predict(cf, as.matrix(test_x), estimate.variance = TRUE)
```

The resulting estimates of observation-level spillover effects can then be visualized and interpreted as necessary for one's purposes. In the subsequent example of spillover effects in anti-bullying programs, I aggregate observation-level effect estimates within schools and then compare the school-level distribution of estimates for a set of four models, each mod-

[12]This code is drawn from https://www.markhw.com/blog/causalforestintro, where additional useful code is available for using a causal forest in a social sciences context.

[13]If all observations are plotted together at once, then the standard errors can produce a messy visual which is difficult to interpret. But if one is interested in a single or handful of cases, then the standard errors about predicted observation-level causal effects, conditional on covariates, are incredibly helpful for probabilistically comparing effect estimates.

eling a different outcome.

## 3.4 Simulation

In observational studies, heterogeneous treatment effects are estimated based on the observed outcomes for each unit and their covariate values, through which moderating effects can occur. In this section I simulate data where spillover effects occur through indirect treatment exposure and all potential outcomes are generated for each observation. Each observation's potential outcomes $(d_{11}, d_{10}, d_{01}, d_{00})$ are simulated to be a function of: whether or not it receives direct treatment, it is tied to a treated unit, and its covariate values. Spillovers are then programmed in to explicitly occur heterogeneously, conditional on covariate values. The two variables that I vary throughout the simulation are 1) the number of observations and 2) the heterogeneity of the true spillover effects. The simulations therefore tests the ability of the combination of causal forests and estimated spillover exposure weights to correctly estimate a unit's spillover effect, keeping the number of observations and effect heterogeneity in mind.

Since the causal forest's final product is a set of observation-level estimates of causal effects, it follows to simulate observations where potential outcomes are known and to evaluate how closely the causal forest estimates the relevant differences between the potential outcomes of interest, relative to the true observation-level difference. Alongside evaluating whether or not the proposed procedure returns accurate estimates of the true causal effect, I compare the final estimate accuracies when the inverse probability weights are included and not included, testing whether weighting observations by their probability of receiving indirect treatment exposure improves the causal forest's accuracy.

This simulation's goal is to demonstrate that the causal forest can be used to model heterogeneous spillover effects and that the causal forest effectively implements the inverse probability weights. While forest-based methods are confirmed to hold promise for heterogeneous treatment effects and the inverse probability weights are recognized to assist

in spillover studies, I hope to convince the reader that the combination of the approaches works as expected. Across simulations, the causal forest returns accurate unit-level estimates of spillover effects. Moreover, while implementing inverse probability weights almost never produces less accurate estimates, it generally improves or does not change the model's accuracy. In game-theoretic terms, including the probability weights is a 'weakly-dominant strategy'; including the weights almost never leaves the researcher worse off, but generally helps.

### 3.4.1 Simulation Setup

Each simulation iteration is an experiment with sample size $n$, a network of ties between units, and treatment assignment $z$. Our effect of interest is the spillover effect, where treatment is randomly assigned and then units with a network tie to treated units receive indirect treatment exposure. Our two exposure conditions of interest, then, are indirect exposure only ($d_{01}$) and no exposure ($d_{00}$). This provides the following function for an average spillover effect:

$$\tau = \mathbb{E}[y(d_{01}) - y(d_{00})] \tag{3.8}$$

which is explicitly programmed to vary across covariate values, creating the conditional average spillover effect:

$$\tau(x) = \mathbb{E}[y(d_{01}) - y(d_{00}) \mid X = x]. \tag{3.9}$$

The benefit of simulation is that we are able to know every observation's entire set of potential outcomes. While the estimator faces the fundamental problem of causal inference, only encountering one potential outcome for each observation, we are able to verify the causal forest's observation-level causal effect estimates through the known potential outcomes. Each observation is assigned 5 covariate values, $X$, drawn from a standard normal distribution, and a set of 5 spillover effects for those covariates are drawn from a normal distribution with mean 0 and varying standard deviation (one of our varying parameters), which captures the true underlying effect heterogeneity. To put this in familiar terms, let

the spillover coefficients be labelled $\beta$. The true direct treatment effect is held arbitrarily at 2. Each unit then receives some random noise that cannot be modelled, $\varepsilon_i$. Each unit's potential outcomes are simulated as follows:

- Direct and indirect exposure ($d_{11}$): $2 + X_i * \beta + \varepsilon_i$
- Direct exposure only ($d_{10}$): $2 + \varepsilon_i$
- Indirect exposure only ($d_{01}$): $X_i * \beta + \varepsilon_i$
- No exposure ($d_{00}$): $\varepsilon_i$

For each unit, every potential outcome is calculated so that the true effects can be compared against the causal forest's estimates. The simulated experiment's treatment assignment and network ties are also held constant within each simulation iteration. Once the probabilities for each condition are estimated, $\boldsymbol{\pi}_i = (\pi_i(d_{11}), \pi_i(d_{10}), \pi_i(d_{01}), \pi_i(d_{00}))$, a causal forest is fit to compare observations with indirect exposure only to observations with no exposure, conditional on their covariate values. The resulting model provides spillover effect estimates for each observation, which are produced by comparing the average outcome for observations with similar covariate values (fall in the same random forest leaf) that receive indirect exposure and those that do not. We can then see how well the causal forest's difference-based estimates map onto the true known observation-level effects.

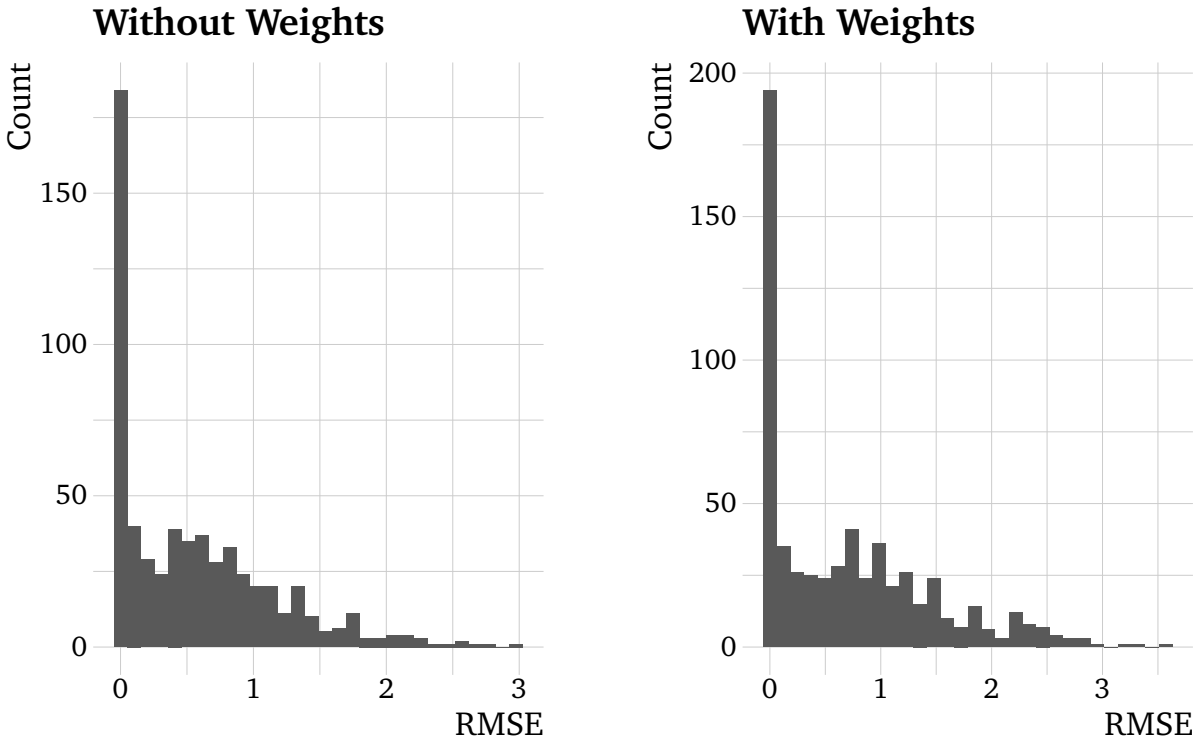
Lastly, across simulations I vary two parameters: the number of observations and the heterogeneity of effects ($\sigma$ in the distribution that $\beta$ is drawn from). For each simulation I also fit a causal forest without the estimated inverse probability of treatment weights and then a causal forest with the weights, allowing a comparison of whether the Horovitz-Thompson weights improve the model's accuracy.

### 3.4.2 Simulation Results

Figures 3.2 and 3.3 include simulation results. Across simulations the causal forests without weights and with weights both tend to correctly estimate the true observation-level spillover effect. This is reflected in Figure 2, where the x-axis is a model's root-mean squared

error for a single iteration, which is calculated by comparing the observation-level estimated spillover effects to the true effects. In Figure 3.2, both the weighted and unweighted models disproportionately tend to accurately estimate the true causal effect, with the histograms heavily peaked at zero. However, foreshadowing Figure 3.3, the peak at zero is higher for the weighted model. Therefore, we see that the causal forest is generally effective at estimating heterogeneous spillover effects, but tends to benefit meaningfully from the inverse probability weights.

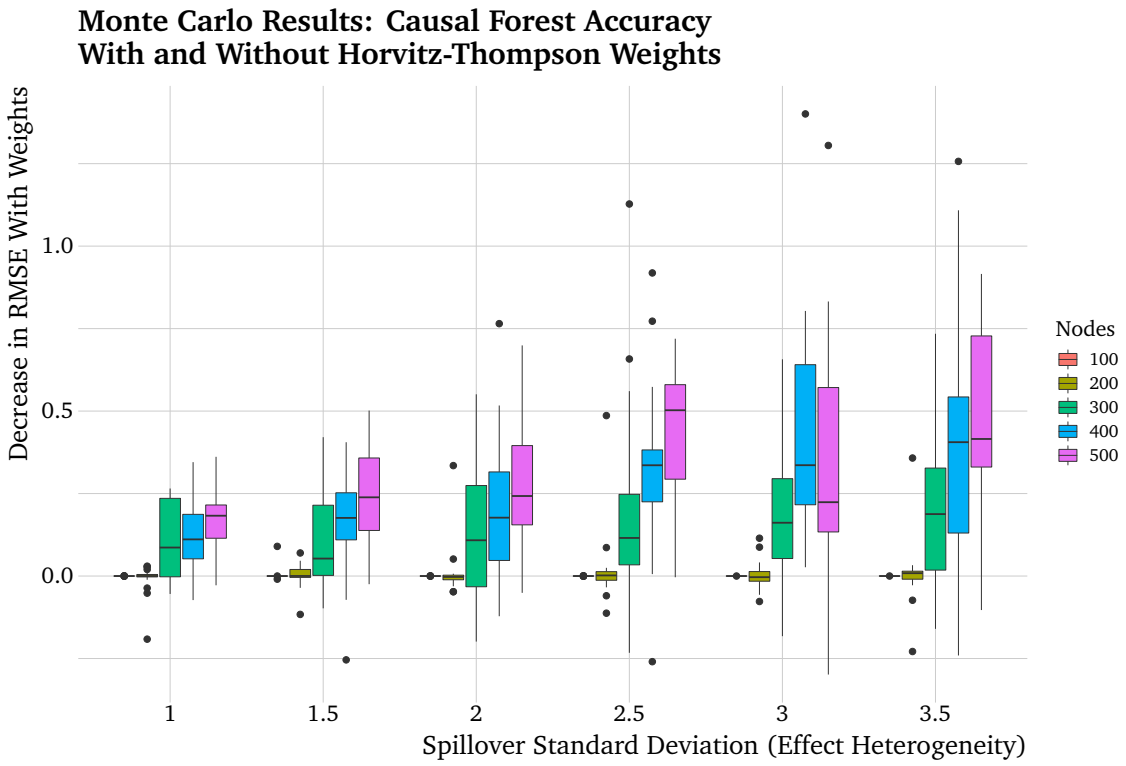
Figure 3.2: Histogram of Model Accuracy With and Without Probability Weights




Histograms of model accuracy in estimating observation-level spillover effects across simulations. The x-axis is a model's root mean squared error (RMSE) and the y-axis is the number of observations where this RMSE occurs. The left-hand histogram is for a causal forest that does not include inverse-probability weights and the right-hand histogram is for a causal forest with the probability weights. Both causal forests are peaked at zero, meaning they tend to return accurate estimates, but the peak is higher at zero for the causal forest with probability weights.

Decomposing the added benefit of the probability weights, Figure 3.3 compares the RMSE for the weighted and unweighted models in each simulation. Positive y-axis values represent a decrease in RMSE when using weights, or an increase in model accuracy. X-axis values represent the standard deviation in spillover effects – where a larger standard deviation represents more heterogeneous effects. The color of each boxplot is the number of observations in the network. We see that as the number of observations increases, so too does the value of the inverse probability weights. This is likely due to the fact that the more observations there are, then the more distinct the probability of being in each exposure condition becomes, increasing the impact of the weights. In addition, the more heterogeneous the underlying effects are, then the more these weights tend to increase model accuracy for larger networks.

Figure 3.3: Causal Forest RMSE By Sample Size and Effect Heterogeneity




RMSE for a weighted causal forest minus the RMSE for an unweighted causal forest. The heterogeneity of the true spillover effects increases along the x-axis and the number of observations per simulation is color-coded.

Reiterating the aforementioned point about including weights being akin to a weakly-dominant strategy in game theory, only in a very small handful of simulations does including weights decrease the causal forest's accuracy. Moreover, if the network includes fewer than 200 nodes, then the weights tend to make no difference. However, as effects become more heterogeneous and the network's size grows, then the probability weights tend to increasingly improve model accuracy, relative to an unweighted model.

## 3.5 Application: Anti-Bullying Programs in School

Turning to a demonstration of the method in an applied setting, Paluck et al. (2016) provide an example of quantifying spillover effects in a randomized controlled trial. In their study, Paluck et al. measure social networks within schools in New Jersey and then assign prominent nodes – popular students – at random to an anti-bullying program. The study demonstrates the effectiveness of anti-bullying programs in schools, not just comparing subsequent aggregate behavior at the school level, but also comparing outcomes of students that are friends with students in the anti-bullying program to students who are not friends with any attendees – spillover effects. On average, Paluck et al. find that the anti-bullying programs, though implemented with a small number of students, tend to decrease bullying at the school-level and behaviors differ meaningfully between those with indirect exposure through friends and those with no exposure.

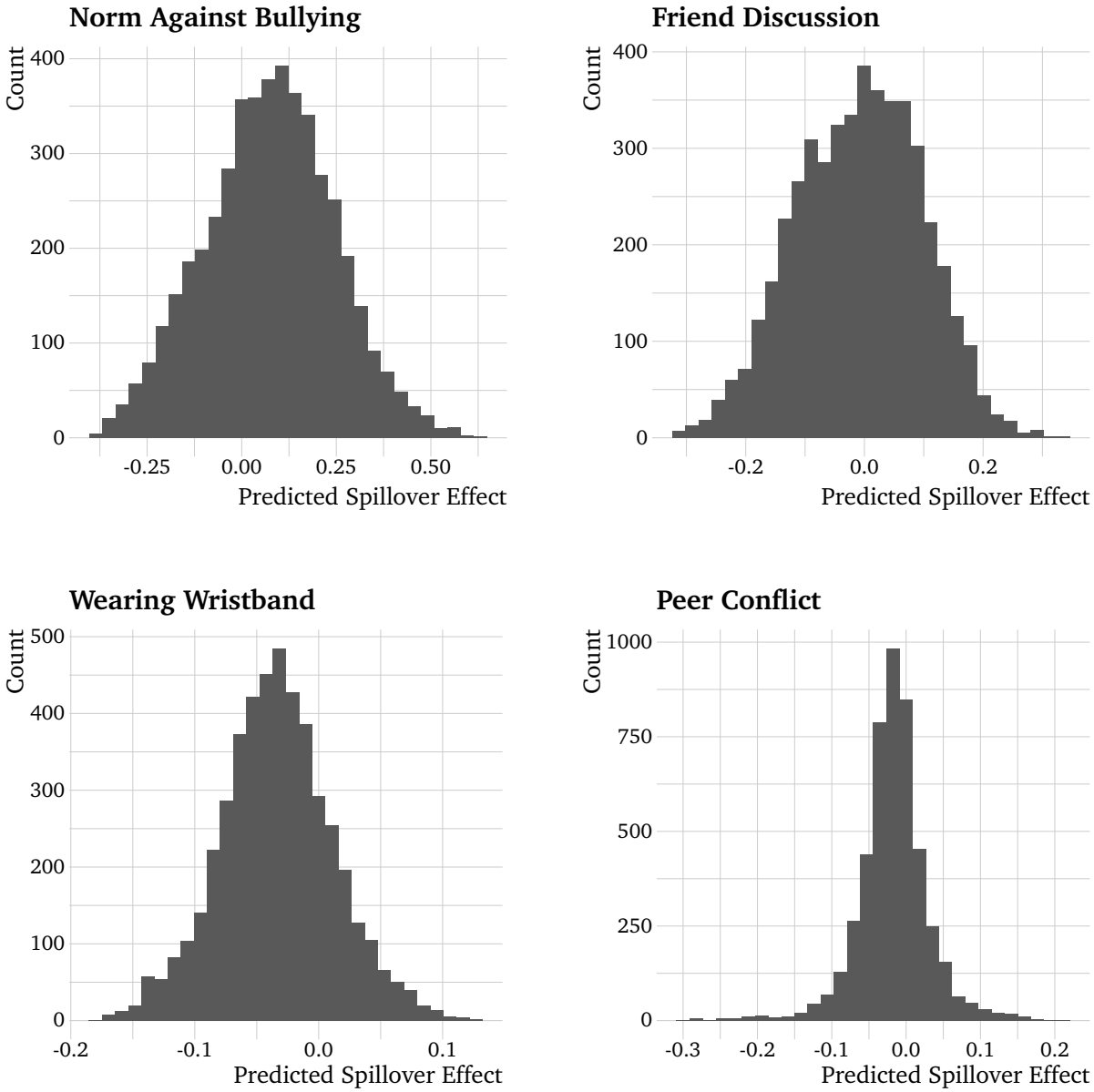
The study is normatively important and methodologically thorough. It also includes all necessary pieces for examining the degree to which the anti-bullying program's spillover effects also vary across groups and are heterogeneous. In this section I extend the paper's results, estimating each student's probability weights for each exposure condition and then fitting a causal forest for each of the four dependent variables included in the study. The four dependent variables are: 1) a perceived norm against bullying within the school, 2) frequency of discussions about bullying among friends, 3) wearing a wristband to signify support for the anti-bullying program, and 4) subsequent conflictual behavior between

peers at the school (based on administrative records). Evidence of an effective anti-bullying program is a positive association with 1-3 and a negative association with 4 (a decrease in subsequent conflict between peers). Covariates included in the causal forest are: a numeric school ID, the student's grade level, ethnicity, gender, proxies for income, and the student's recent disciplinary record.

I fit a causal forest with data on students in treated schools, comparing students in the control group (not assigned to the anti-bullying program) who are friends with students in the program to control group students not friends with students in the program. Another form of spillover effects, which I briefly examine here, is to compare treated students who are and are not friends with other treated students. Figure 4 includes histograms with the distribution of predicted spillover effects for each student, conditional on their covariate values. For these dependent variables, wearing a wristband is a binary outcome, whereas the others are 3 or 4 point scales. This means the spillover effects are of a relatively small magnitude, but that is reasonable given the size of the study and stickiness of social behaviors in schools. However, we do see heterogeneity in the direction of the spillover effects, which is an important distinction. Across all outcomes, a substantial portion of students fall under both positive and negative spillover effects. This suggests that if the program were reimplemented at a larger scale, we should expect that, in terms of spillover effects, even if the program is more effective than not, in some schools the program would be likely to backfire.

Turning to how spillover effects vary across schools, I cluster students by schools and compare the within-school distributions of predicted spillover effects. As an aside, these results are not identical to running a regression with interaction effects for each school. Rather, Figure 5 takes the causal forest's estimates, which are a function of each observation's entire set of covariate values, and then groups the estimates by schools. So while the output may appear as if only outcomes and school ID's are considered in model fitting, the results actually reflect the entirety of the covariates for each observation. Looking at the within-school estimates for each outcome, we see where the effects vary in magnitude

Figure 3.4: Distribution of Spillover Effect Estimates by Outcome

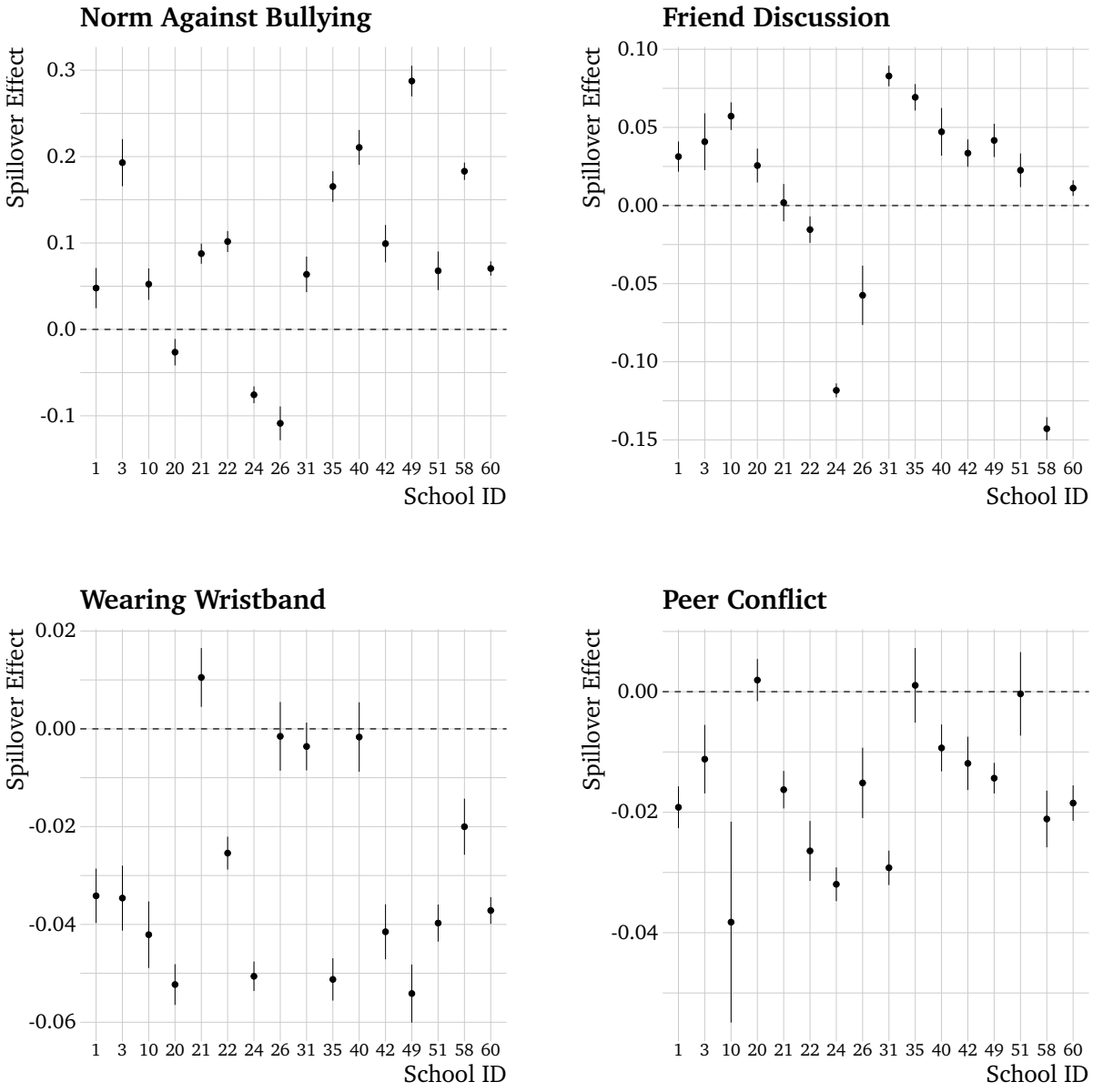



Histogram of estimated spillover effects for each dependent variable. While the dataset includes schools where there was no anti-bullying program, I only consider students in schools with a program here.

and direction. Without more information about the schools it is difficult to say why we see the variation we do across these schools, but we can see that some schools are far more receptive than others to the anti-bullying program, with a handful of schools having the opposite effect. The heterogeneity of these results demonstrates how taking evidence of

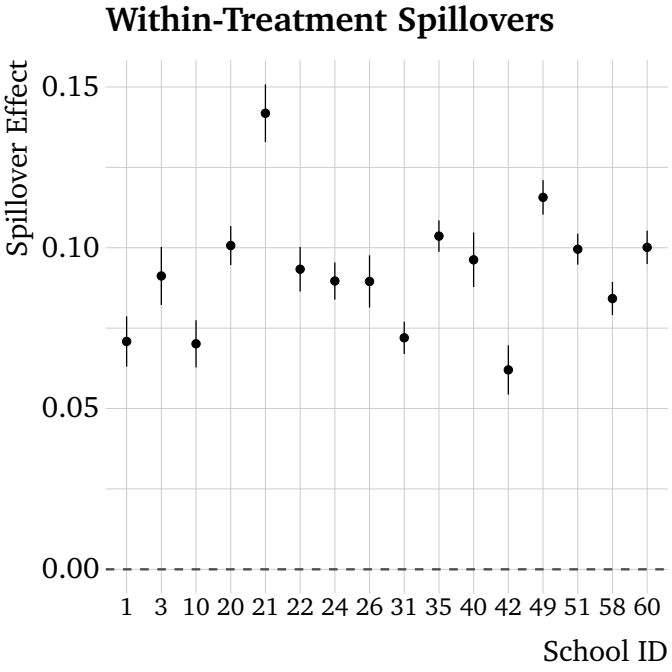
generally effective spillovers can be dangerous and misleading in certain contexts because the effects are heterogeneous across groups.

Figure 3.5: Spillover Effects by Dependent Variable and School




Spillover effects across dependent variables clustered by schools. Effects are calculated at the individual level and then grouped and summarized for each school. Uncertainty is reflected in the distribution of point-estimates for each individual within each school.

Figure 3.6: Spillover Effect In Treatment Group




Spillover effects for wearing the anti-conflict wristband when observations in the treatment group that share a tie other treated units are compared to observations in the treatment group without a tie to treated units.

Curiously, while three out of the four dependent variables demonstrate spillovers associated with the desired outcome (more of a norm against bullying, more discussions about bullying among friends, and less subsequent conflicts), we see the opposite with wearing wristbands. Instead, the causal forest estimates that the spillover effects are generally negative – causing less wristband wearing than would otherwise be the case. This clashes with the expectation from the original paper, where the anti-bullying program is associated with wearing more wristbands across the schools. In response, in Figure 6, I compare treated students with a tie to other treated students to treated students with no indirect exposure – $\tau(d_{11}, d_{10})$ – and see consistently large spillover effects. This suggests that the results around wristband wearing are driven largely by the popular students selected into the anti-bullying program and not the other students. This spillover effect is generally larger and

in the expected direction, suggesting that not all forms of spillovers work the same way, another interesting topic for discussion.

## 3.6 Conclusion

This article develops a procedure for estimating heterogeneous spillover effects. The proposed methodology combines techniques for estimating spillover effects with causal random forests. Each observation's potential outcomes are decomposed to not only differentiate by treatment status, but also whether an observation shares a network tie with a treated unit. Then weights are estimated for the probability that each observation ends up in each of the decomposed potential outcomes. Lastly, the weights are included in a causal random forest that is built to flexibly model heterogeneous treatment effects, but instead of comparing observations by their treatment status, observations are compared based on whether they share a network tie with treated units. After verifying the method's efficacy in simulations, I apply it to a study of spillover effects in anti-bullying programs at schools. I find that certain schools do tend to respond favorably, whereas others do not. This example shows how a pilot study, like the anti-bullying program, can inform where subsequent efforts can be targeted to maximize their effectiveness.

# Bibliography


G. Allison. *Destined for War: Can America and China Escape Thucydides's Trap?* Houghton Mifflin Harcourt, 2017.

C. C. Anderson, S. M. Mitchell, and E. U. Schilling. Kantian dynamics revisited: Time-varying analyses of dyadic igo-conflict relationships. *International Interactions*, 42(4): 644–676, 2016.

P. M. Aronow, C. Samii, et al. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4): 1912–1947, 2017.

P. M. Aronow, D. Eckles, C. Samii, and S. Zonszein. Spillover effects in experimental data. *arXiv preprint arXiv:2001.05444*, 2020.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

K. Baicker. The spillover effects of state spending. *Journal of public economics*, 89(2-3): 529–544, 2005.

M. A. Bailey, A. Strezhnev, and E. Voeten. Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution*, 61(2):430–456, 2017.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

K. Barbieri and O. Keshk. Correlates of war project trade data set codebook, version 4.0. *Online: http://correlatesofwar.org*, 2016.

K. Barbieri, O. M. Keshk, and B. M. Pollins. Trading data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science*, 26(5):471–491, 2009.

R. Bayer. Diplomatic exchange data set, v2006. 1. *Online: http://correlatesofwar. org*, 2006.

M. Beckley. China's century? why america's edge will endure. *International Security*, 36 (3):41–78, 2012.

M. Beckley. *Unrivaled: Why America will remain the world's sole superpower*. Cornell University Press, 2018.

T. Brambor, W. R. Clark, and M. Golder. Understanding interaction models: Improving empirical analyses. *Political analysis*, 14(1):63–82, 2006.

B. F. Braumoeller. Hypothesis testing and multiplicative interaction terms. *International organization*, 58(4):807–820, 2004.

B. F. Braumoeller. *The great powers and the international system: systemic theory in empirical perspective*. Cambridge University Press, 2013.

B. F. Braumoeller. *Only the dead: the persistence of war in the modern age*. Oxford University Press, 2019.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

L. Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001b.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

R. S. Burt et al. *Brokerage and closure: An introduction to social capital*. Oxford university press, 2005.

E. H. Carr. The twenty years' crisis, 1919-1939: an introduction to the study of international relations. 1946.

R. J. Carroll and B. Kenkel. Prediction, proxies, and power. *American Journal of Political Science*, 2016.

S. Chan. Can't get no satisfaction? the recognition of revisionist states. *International relations of the Asia-pacific*, 4(2):207–238, 2004.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

K.-y. Cheung and L. Ping. Spillover effects of fdi on innovation in china: Evidence from the provincial data. *China economic review*, 15(1):25–44, 2004.

T. J. Christensen. Fostering stability or creating a monster? the rise of china and us policy toward east asia. *International security*, 31(1):81–126, 2006.

A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

S. J. Cranmer and B. A. Desmarais. A critique of dyadic design. *International Studies Quarterly*, 60(2):355–362, 2016.

S. J. Cranmer and B. A. Desmarais. What can we learn from predictive modeling? *Political Analysis*, 25(2):145–166, 2017.

S. J. Cranmer, P. Leifeld, S. D. McClurg, and M. Rolfe. Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*, 61(1):237–251, 2017.

A. Dafoe, J. Renshon, and P. Huth. Reputation and status as motives for war. *Annual Review of Political Science*, 17:371–393, 2014.

J. Davidson. *The origins of revisionist and status-quo states*. Springer, 2006.

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Y. Deng. *China's struggle for status: the realignment of international relations*. Cambridge University Press, 2008.

M. Duffy Toft. Population shifts and civil war: A test of power transition theory. *International Interactions*, 33(3):243–269, 2007.

M. G. Duque. Recognizing international status: A relational approach. *International Studies Quarterly*, 2018.

D. M. Edelstein. *Over the Horizon: Time, Uncertainty, and the Rise of Great Powers*. Cornell University Press, 2017.

M. Emirbayer. Manifesto for a relational sociology. *American journal of sociology*, 103(2): 281–317, 1997.

M. Emirbayer and J. Goodwin. Network analysis, culture, and the problem of agency. *American journal of sociology*, 99(6):1411–1454, 1994.

E. Erikson. Formalist and relationalist theory in social network analysis. *Sociological Theory*, 31(3):219–242, 2013.

J. Esarey and J. L. Sumner. Marginal effects in interaction models: Determining and controlling the false positive rate. *Comparative Political Studies*, 51(9):1144–1176, 2018.

J. D. Fearon. Rationalist explanations for war. *International organization*, 49(3):379–414, 1995.

M. Finnemore and K. Sikkink. International norm dynamics and political change. *International organization*, 52(4):887–917, 1998.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38 (4):367–378, 2002.

R. Friedman Lissner and M. Rapp-Hooper. The day after trump: American strategy for a new international order. *The Washington Quarterly*, 41(1):7–25, 2018.

D. M. Gibler, S. V. Miller, and E. K. Little. An analysis of the militarized interstate dispute (mid) dataset, 1816–2001. *International Studies Quarterly*, 60(4):719–730, 2016.

R. Gilpin. *War and change in world politics*. Cambridge University Press, 1983.

S. E. Goddard. When right makes might: how prussia overturned the european balance of power. *International Security*, 33(3):110–142, 2009.

S. E. Goddard. *When Right Makes Might: Rising Powers and World Order*. Cornell University Press, 2018a.

S. E. Goddard. Embedded revisionism: Networks, institutions, and challenges to world order. *International Organization*, pages 1–35, 2018b.

E. Goh. Meeting the china challenge: The us in southeast asian regional security strategies. 2005.

E. Goh. *The struggle for order: Hegemony, hierarchy, and transition in post-Cold War East Asia*. Oxford University Press, 2013.

D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

B. M. Greenwell. pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1):421–436, 2017. doi: 10.32614/RJ-2017-016. URL https://doi.org/10.32614/RJ-2017-016.

J. Grimmer, S. Messing, and S. J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

E. M. Hafner-Burton, M. Kahler, and A. H. Montgomery. Network analysis for international relations. *International Organization*, 63(3):559–592, 2009.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

J. Haushofer and J. Shapiro. The long-term impact of unconditional cash transfers: experimental evidence from kenya. *Busara Center for Behavioral Economics, Nairobi, Kenya*, 2018.

J. Hill and Y.-S. Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

J. Hill, A. Linero, and J. Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7, 2020.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

O. R. Holsti, R. M. Siverson, and A. L. George. *Change in the international system*. Routledge, 2019.

M. C. Horowitz, A. C. Stam, and C. M. Ellis. *Why leaders fight*. Cambridge University Press, 2015.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860, 2008.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

P. T. Jackson and D. H. Nexon. Relations before states: Substance, process and the study of world politics. *European journal of international relations*, 5(3):291–332, 1999.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

L. R. James and J. M. Brett. Mediators, moderators, and tests for mediation. *Journal of applied psychology*, 69(2):307, 1984.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

L. Jenke and C. Gelpi. Theme and variations: Historical contingencies in the causal model of interstate conflict. *Journal of Conflict Resolution*, 61(10):2262–2284, 2017.

R. Jervis. *System effects: Complexity in political and social life*. Princeton University Press, 1998.

A. I. Johnston. Treating international institutions as social environments. *International Studies Quarterly*, 45(4):487–515, 2001.

A. I. Johnston. Is china a status quo power? *International security*, 27(4):5–56, 2003.

A. I. Johnston and R. S. Ross. *Engaging China: The management of an emerging power*, volume 24. Psychology Press, 1999.

J. J. Jones, R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler. Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 us presidential election. *PloS one*, 12(4), 2017.

K. Kadera. *The power-conflict story: A dynamic model of interstate rivalry*. University of Michigan Press, 2001.

A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

P. Kennedy. *The rise and fall of the great powers: economic change and military conflict from 1500 to 2000*. Vintage, 2010.

D. Kent, J. D. Wilson, and S. J. Cranmer. A randomization approach to dynamic analysis of panel data. *Conditionally Accepted at Political Analysis*, 2020.

J. D. Kertzer. *Resolve in international politics*. Princeton University Press, 2016.

G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

D. A. Lake. *Hierarchy in international relations*. Cornell University Press, 2009.

D. W. Larson and A. Shevchenko. Status seekers: Chinese and russian responses to us primacy. *International Security*, 34(4):63–95, 2010.

G. Lawson. Revolutions and the international. *Theory and Society*, 44(4):299–319, 2015.

R. N. Lebow and B. Valentino. Lost in transition: A critical analysis of power transition theory. *International Relations*, 23(3):389–410, 2009.

B. Leeds, J. Ritter, S. Mitchell, and A. Long. Alliance treaty obligations and provisions, 1815-1944. *International Interactions*, 28(3):237–260, 2002.

D. Lemke and W. Reed. Regime types and status quo evaluations: Power transition theory and the democratic peace. *International Interactions*, 22(2):143–164, 1996.

P. Y. Lipscy. *Renegotiating the World Order: Institutional Change in International Relations*. Cambridge University Press, 2017.

J. M. Lyall. *Paths of Ruin: Why Revisionist States Arise and Die in World Politics*. Cornell University, 2005.

P. K. MacDonald. *Networks of Domination: The Social Foundations of Peripheral Conquest in International Politics*. OUP Us, 2014.

P. K. MacDonald. Embedded authority: a relational network approach to hierarchy in world politics. *Review of International Studies*, 44(1):128–150, 2018.

P. K. MacDonald and J. M. Parent. *Twilight of the Titans: Great Power Decline and Retrenchment*. Cornell University Press, 2018.

M. G. Marshall, T. R. Gurr, C. Davenport, and K. Jaggers. Polity iv, 1800-1999: Comments on munck and verkuilen. *Comparative Political Studies*, 35(1):40–45, 2002.

R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

J. J. Mearsheimer. Can china rise peacefully? *The National Interest*, 25:23–37, 2014.

S. Minhas, P. D. Hoff, and M. D. Ward. Inferential approaches for network analysis: Amen for latent factor models. *Political Analysis*, 27(2):208–222, 2019.

E. B. Montgomery. *In the Hegemon's Shadow: Leading States and the Rise of Regional Powers*. Cornell University Press, 2016.

H. Morgenthau. Politics among nations: The struggle for power and peace. *Nova York, Alfred Kopf*, 1948.

A. I. Naimi and L. B. Balzer. Stacked generalization: an introduction to super learning. *European journal of epidemiology*, 33(5):459–464, 2018.

M. Neunhoeffer and S. Sternberg. How cross-validation can go wrong and what to do about it. *Political Analysis*, 27(1):101–106, 2019.

A. Ng. Volatility spillover effects from japan and the us to the pacific–basin. *Journal of international money and finance*, 19(2):207–233, 2000.

D. W. Nickerson. Is voting contagious? evidence from two field experiments. *American political Science review*, 102(1):49–57, 2008.

A. F. Organski and J. Kugler. *The war ledger*. University of Chicago Press, 1981.

G. Palmer, V. D'Orazio, M. R. Kenwick, and R. W. McManus. Updating the militarized interstate dispute data: A response to gibler, miller, and little. *International Studies Quarterly*, 2019.

E. L. Paluck, H. Shepherd, and P. M. Aronow. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571, 2016.

T. Paul. *Accommodating rising powers: past, present, and future*. Cambridge University Press, 2016.

T. V. Paul, D. W. Larson, and W. C. Wohlforth. *Status in world politics*. Cambridge University Press, 2014.

E. C. Polley and M. J. Van Der Laan. Super learner in prediction. 2010.

Y. Qin. A relational theory of world politics. *International Studies Review*, 18(1):33–47, 2016.

J. Radford and K. Joseph. Theory in, theory out: The uses of social theory in machine learning for social science. *arXiv*, pages arXiv–2001, 2020.

J. Renshon. Status deficits and war. *International Organization*, 70(3):513–550, 2016.

J. Renshon. *Fighting for status: hierarchy and conflict in world politics*. Princeton University Press, 2017.

D. B. Rubin. Formal models of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990.

K. Schake. *Safe Passage: The Transition from British to American Hegemony*. Harvard University Press, 2017.

P. W. Schroeder. *The transformation of European politics, 1763-1848*. Oxford University Press, 1994.

R. L. Schweller. Bandwagoning for profit: Bringing the revisionist state back in. *International Security*, 19(1):72–107, 1994.

R. L. Schweller. Managing the rise of great powers: history and theory. *Engaging China: The management of an emerging power*, pages 1–31, 1999.

R. L. Schweller. Rising powers and revisionism in emerging international orders. *Russia in Global Affairs*, 7, 2015.

J. R. I. Shifrinson. *Rising Titans, Falling Giants: How Great Powers Exploit Power Shifts*. Cornell University Press, 2018.

G. Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

J. D. Singer, S. Bremer, J. Stuckey, et al. Capability distribution, uncertainty, and major power war, 1820-1965. *Peace, war, and numbers*, 19:48, 1972.

D. Snidal. Relative gains and the pattern of international cooperation. *American Political Science Review*, 85(3):701–726, 1991.

J. Snyder. Civil-military relations and the cult of the offensive, 1914 and 1984. *International Security*, 9(1):108–146, 1984.

J. Snyder. *Myths of empire: Domestic politics and international ambition*. Cornell University Press, 1991.

Stockholm International Peace Research Institute. *SIPRI Yearbook 2019: Armaments, Disarmament and International Security*. Oxford University Press, 2019.

S. J. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. In *Complex Spreading Phenomena in Social Systems*, pages 289–322. Springer, 2018.

The H2O.ai team. *h2o: R Interface for H2O*, 2015. URL http://www.h2o.ai. R package version 3.1.0.99999.

C. G. Thies and M. D. Nieman. *Rising Powers and Foreign Policy Revisionism: Understanding BRICS Identity and Behavior Through Time*. University of Michigan Press, 2017.

D. H. Tingley. The dark side of the future: An experimental test of commitment problems in bargaining. *International Studies Quarterly*, 55(2):521–544, 2011.

M. Trachtenberg. Audience costs: An historical analysis. *Security Studies*, 21(1):3–42, 2012.

A. Truong, A. Walters, J. Goodsitt, K. Hines, B. Bruss, and R. Farivar. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. *arXiv preprint arXiv:1908.05557*, 2019.

J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

S. Van Evera. The cult of the offensive and the origins of the first world war. *International Security*, 9(1):58–107, 1984.

T. J. VanderWeele and E. J. T. Tchetgen. Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & probability letters*, 81(7):861–869, 2011.

J. N. Victor, A. H. Montgomery, and M. Lubell. *The Oxford Handbook of Political Networks*. Oxford University Press, 2017.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

S. M. Walt. *Revolution and war*. Cornell University Press, 1996.

S. Ward. *Status and the Challenge of Rising Powers*. Cambridge University Press, 2017.

S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

J. L. Weeks. Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1):35–64, 2008.

J. D. Wilson, M. J. Denny, S. Bhamidi, S. J. Cranmer, and B. A. Desmarais. Stochastic weighted graphs: Flexible model specification and simulation. *Social Networks*, 49:37–47, 2017.

R. Wolf. Respect and disrespect in international politics: the significance of status recognition. *International Theory*, 3(1):105–142, 2011.

A. Wolfers. Discord and collaboration: Essays on international politics, 1962.

# Appendix A: Appendix: Chapter 1

Each of the following table references a time period in international politics and ranks the top ten average expected benefit scores. Each states's average CINC score and average win probability across all dyads is included also. The average expected benefit score represents how much of the international sphere a state expects it should have control over, based on power calculations alone. While the traditional great powers tend to be included in the tables, rankings and percentages vary substantially over time.

Table A.1: Top Pre-WWI Capability Estimates

| **Country** | **Average CINC** | **Average Win Probability** | **Average Expected Benefits** |
|---|---|---|---|
| United Kingdom | 0.245 | 0.749 | 0.186 |
| United States | 0.112 | 0.823 | 0.0933 |
| France | 0.113 | 0.714 | 0.0814 |
| Russia | 0.122 | 0.534 | 0.0636 |
| Germany | 0.0847 | 0.630 | 0.0538 |
| China | 0.148 | 0.326 | 0.0467 |
| Austria-Hungary | 0.0621 | 0.590 | 0.363 |
| Spain | 0.0237 | 0.617 | 0.0147 |
| Turkey | 0.0351 | 0.414 | 0.0143 |
| Japan | 0.0264 | 0.490 | 0.0131 |

Table A.2: Top Interwar Period Capability Estimates

| Country | Average CINC | Average Win Probability | Average Expected Benefits |
|---|---|---|---|
| United States | 0.230 | 0.749 | 0.183 |
| Russia | 0.137 | 0.684 | 0.0939 |
| China | 0.128 | 0.537 | 0.0678 |
| United Kingdom | 0.0873 | 0.732 | 0.0639 |
| Germany | 0.0889 | 0.704 | 0.0627 |
| France | 0.0577 | 0.795 | 0.0407 |
| Japan | 0.0042 | 0.632 | 0.0280 |
| Italy | 0.0350 | 0.688 | 0.0241 |
| Poland | 0.0212 | 0.656 | 0.0139 |
| Spain | 0.0164 | 0.649 | 0.0106 |

Table A.3: Top Cold War Capability Estimates

| Country | Average CINC | Average Win Probability | Average Expected Benefits |
|---|---|---|---|
| United States | 0.202 | 0.787 | 0.159 |
| Russia | 0.168 | 0.732 | 0.123 |
| China | 0.111 | 0.663 | 0.0738 |
| India | 0.0517 | 0.648 | 0.0335 |
| Japan | 0.0459 | 0.654 | 0.0301 |
| United Kingdom | 0.0397 | 0.732 | 0.0290 |
| West Germany | 0.0339 | 0.720 | 0.0245 |
| East Germany | 0.0293 | 0.714 | 0.0209 |
| France | 0.0266 | 0.710 | 0.0189 |
| Italy | 0.0191 | 0.689 | 0.0132 |

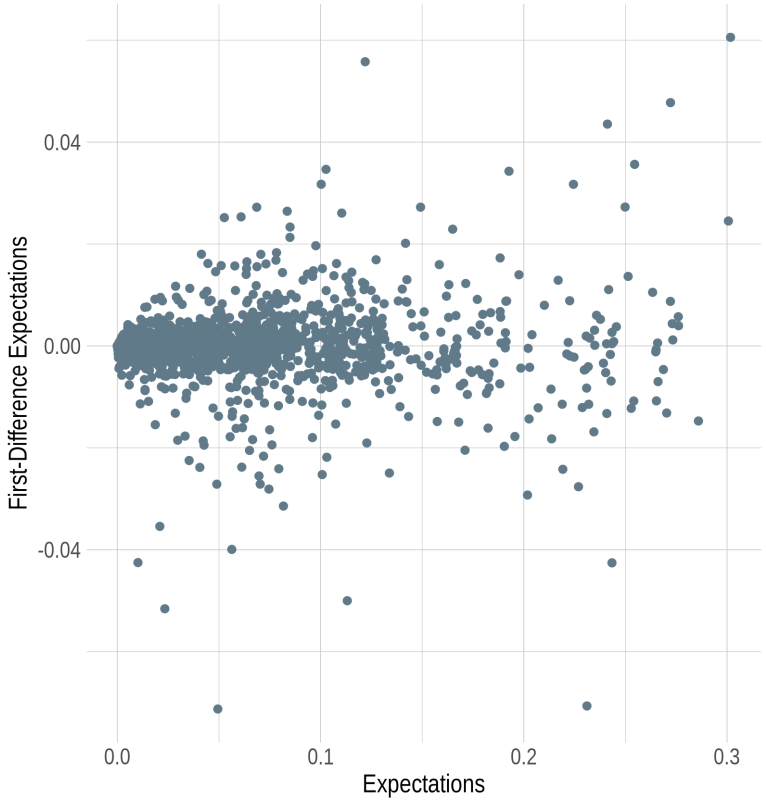
Table A.4: Top Post-Cold War Capability Estimates

| Country | Average CINC | Average Win Probability | Average Expected Benefits |
|---|---|---|---|
| United States | 0.146 | 0.783 | 0.114 |
| China | 0.166 | 0.681 | 0.113 |
| India | 0.0715 | 0.674 | 0.0482 |
| Russia | 0.0524 | 0.718 | 0.0377 |
| Japan | 0.0463 | 0.676 | 0.0313 |
| Germany | 0.0256 | 0.716 | 0.0183 |
| Brazil | 0.0251 | 0.722 | 0.0181 |
| South Korea | 0.0231 | 0.649 | 0.0150 |
| United Kingdom | 0.0210 | 0.706 | 0.0149 |
| France | 0.0196 | 0.698 | 0.0137 |

# Appendix B: Appendix: Chapter 2

## B.1 Growth Rates By Baseline Capabilities

Figure B.1: Growth Rates by Baseline Size



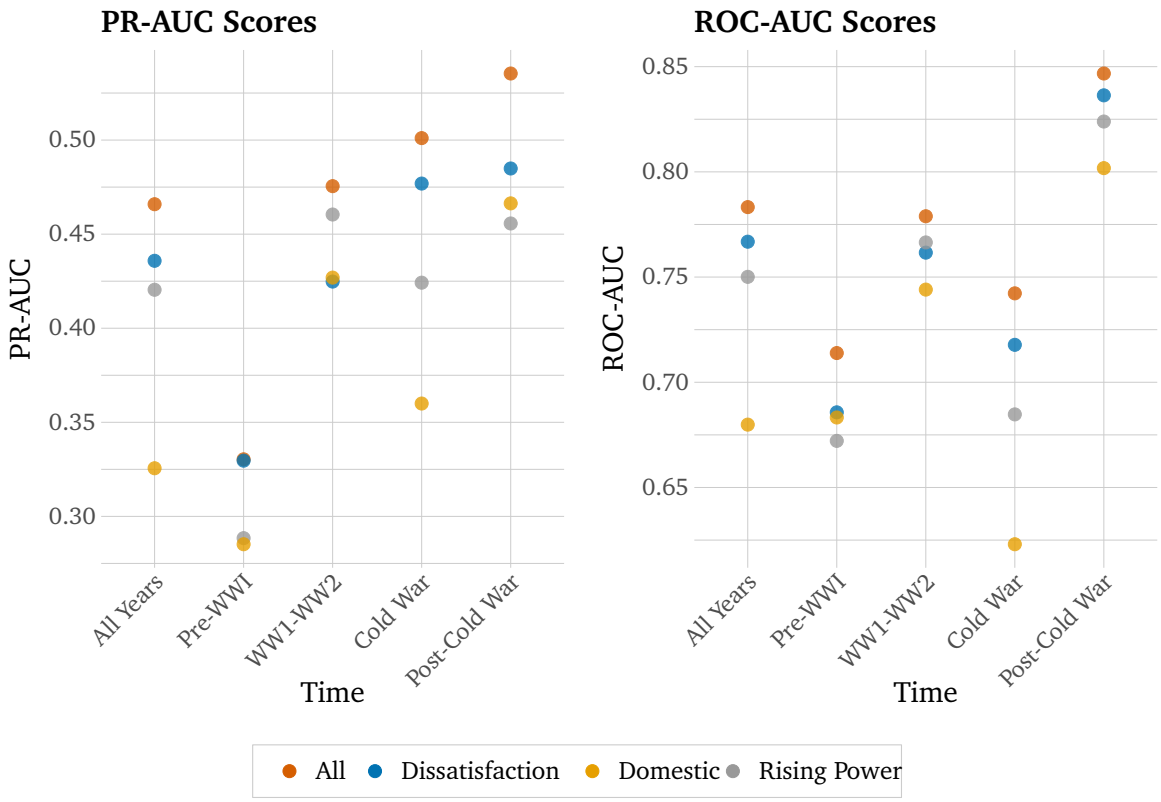

Growth rates over two years compared to baseline state size

This plot demonstrates that the largest growth rates, as operationalized, tend to occur in

the states that are already larger than most. Accordingly, the measure of rising powers does not mistake small growing states that are still unable of revisionist behavior for a genuine revisionist.

## B.2 PR-AUC and ROC-AUC Plots

Figure B.2: Test-Set Accuracy By Time Period and Predictive Variables



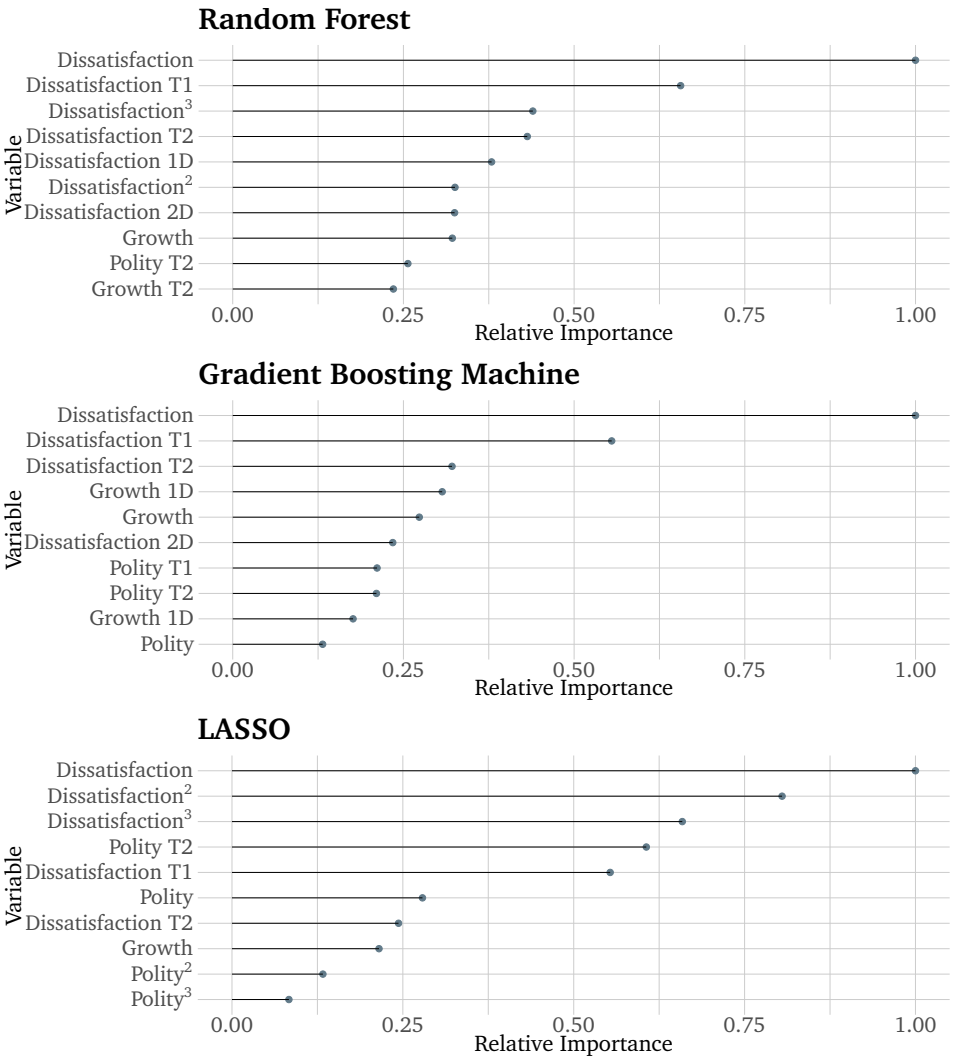

Test set PR-AUC and ROC-AUC for the stacked ensemble across time periods and features included. Higher values on the y-axis represent higher accuracy in test set predictions.

Figure B.2 includes the raw PR-AUC and ROC-AUC scores for the various stacked ensembles fit in Chapter 2. Figure 2.4 solely looks at the comparative difference in AUC scores, relative to the ensemble fit with all possible features, but it does not mention the overall

AUC scores.

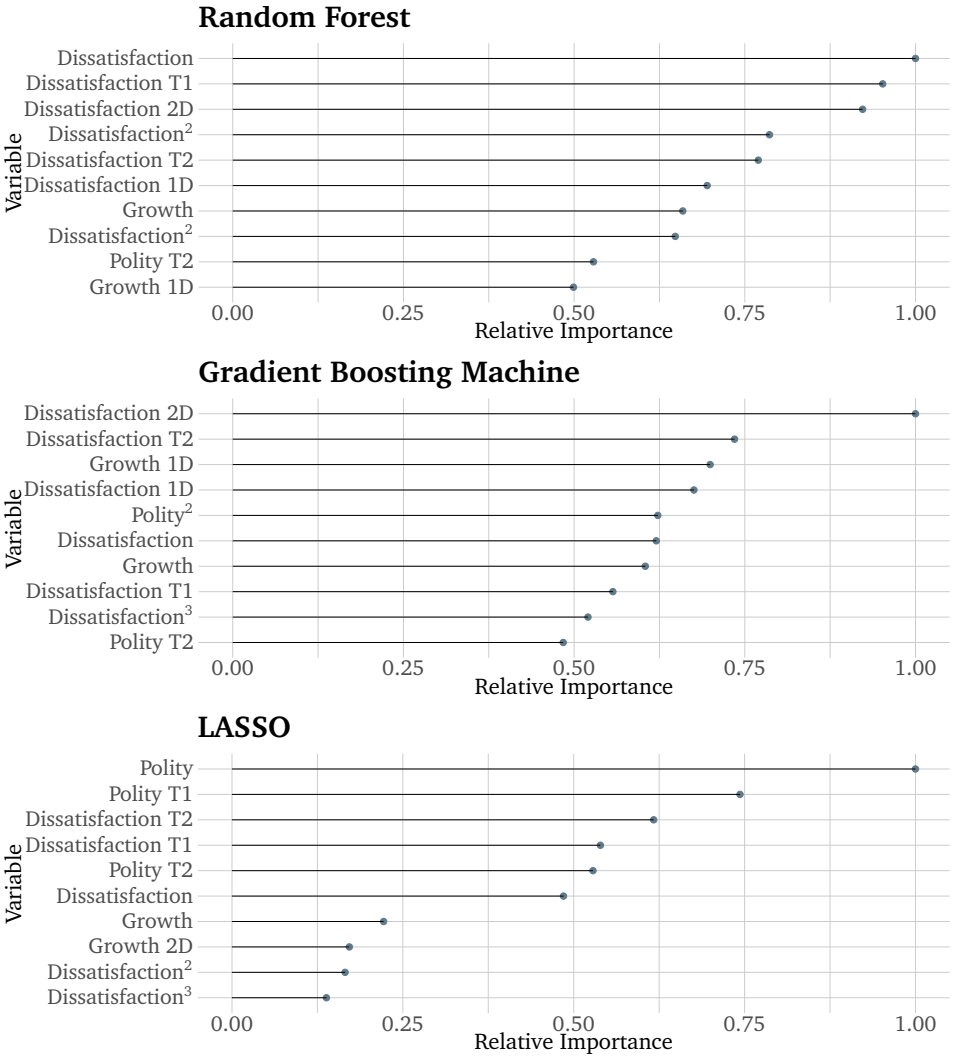
## B.3 Variable Importance Plots Fatal MIDs and Wars

Figure B.3: Variable Importance Plots: Fatal MIDs




Top ten features per model, *in training* when each model is trained on all possible features for all possible years. Relative importance presents a standardized measure of how much a model's loss-function tends to decrease when a variable is included into a model.

Figure B.4: Variable Importance Plots: Wars




Top ten features per model, *in training* when each model is trained on all possible features for all possible years. Relative importance presents a standardized measure of how much a model's loss-function tends to decrease when a variable is included into a model.