Natural Language Processing, Statistical Inference, and American Foreign Policy

Dissertation

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

By

Adam M. Lauretig, B.A., M.A.

Graduate Program in Department of Political Science

The Ohio State University

2019

Dissertation Committee:

William Minozzi, Co-Advisor

Bear F. Braumoeller, Co-Advisor

Brice D.L. Acree

Christopher F. Gelpi

Abstract

What shapes policymakers' foreign policy decisions? Existing work often focuses on how elite decision-making impacts policies and public attitudes, and the role of the public in influencing policy, however, elites themselves have not usually been the object of study in quantitative international relations. This absence is because scholars have lacked both the text of elite deliberations, and the ability to convert this text into data which could be used in statistical inference. In this dissertation, I make contributions on both fronts. To observe elite deliberations, I make use of the recently digitized *Foreign Relations of the United States*, a set of diplomatic documents which capture elite deliberations and decision-making in real time. Using these documents, along with data on casualties and bombing during the Vietnam War, I investigate elite casualty sensitivity, and find that casualties prompt retaliation by elites, in the form of increased bombing. However, casualties have no effect on decisions to end the war, problematizing our understanding of the role of casualty sensitivity in ending wars. I then develop a statistical model, Bayesian Word Embeddings, which builds on existing word embedding models to incorporate measures of uncertainty into the embedding estimates, and then, a way to identify embeddings so they can be used in a regression. I implement this model and method in a `R` package. I use this model to investigate securitization theory, a theory which has largely been studied qualitatively. Investigating securitization across three cases, I find that securitization has no effect

on policy actions by American policymakers. My findings show that incorporating the text of foreign policy decision-making into the study of international relations unsettles and calls into question our previous understanding of states' behaviors in international relations.

For my parents, Susan and Mark,

and

for Anusha, you are my sunshine.

Acknowledgments

This dissertation, like my entire Ph.D. program was not something I could do alone. I benefitted enormously from the help and support of all those around me, and this is my effort to thank everyone who contributed in some way to the completion of this dissertation.

My dissertation would not be possible without the advice and feedback I received from my dissertation committee. My co-chairs Bear Braumoeller and William Minozzi helped me take the raw materials of this dissertation, and craft it into something greater than the sum of its parts. Bear Braumoeller's focus on the international relations implications, and his knowledge of obscure corners of the discipline have helped make this dissertation more intellectually diverse than it otherwise would have been. William Minozzi has been an invaluable mentor and friend throughout graduate school, providing guidance and feedback, talking me through both the big questions: how to do research, how to be a scholar, how to handle the job market, and the smaller questions: how to frame a particular argument, how to work through variational inference, and how to most effectively communicate my findings. In the counterfactual world where he was not my mentor and friend, I would be a much poorer scholar. Discussing word embeddings with Brice Acree has helped sharpen my own thinking what these models are doing, what we can accomplish with them, and how to explain them to a non-technical audience. Chris Gelpi has both provided valuable insights

into the literature on American foreign policy, and had endless insights into what text as data models could be used for.

I have also benefitted from the guidance and mentorship of the faculty in the Department of Political Science at Ohio State. While Jan Pierskalla was not on my committee, I learned a tremendous deal from working with him on two different projects. I learned a great deal about working with massive administrative data, and the process of going from research question to completed paper. Working with Marcus Kurtz has been similarly rewarding, I have enjoyed our conversations, and learned a great deal about causal inference in the presence of time-varying confounders. I would like to thank Jan Box-Steffensmeier for appointing me PRISM fellow in my third year: teaching others about methods has improved my own methodological skills. The second and third chapters of this dissertation have benefitted from the financial support of Michael Neblo and the Institute for Democratic Engagement and Accountability at Ohio State. The rest of the faculty have also provided helpful feedback, and presenting components of this dissertation at the Research in International Politics workshop have helped me clarify my arguments and improve the overall dissertation.

I would be remiss in not thanking the Political Science department staff for all that they do as well. Courtney Sanders has been supremely helpful in navigating the graduate school bureaucracy, and making sure everything is accomplished smoothly. Whenever I have had questions about funding and grant allocations Melodie Mcgrothers has had answers. Robb Hagen has been a great resource when I have had technology questions, and he introduced me to the ASC *Unity* cluster, which made running the models used in this dissertation possible.

The papers in this dissertation were presented at a variety of venues: the Midwest Political Science Association, the Summer Meeting of the Society for Political Methodology, and New Directions in Text as Data. The feedback I received at these conferences improved these papers, and I thoroughly enjoyed the conversations I had with everyone I met at these conferences. The second paper was published in the *Proceedings of the Third Workshop on NLP and Computational Social Science*, and I would like to thank the three reviewers from that conference for their feedback as well.

I owe so much to the faculty at Grinnell College who launched me on this amazing trajectory: Kathy Kamp, Danielle N. Lussier, Wayne Moyer, and John David Stone. Without them, I do not think I would have even considered graduate school.

I would also like to thank my colleagues and friends for the intellectual and emotional support they provided along the way. Conversations with Benjamin Kenzer, Ruthie Pertsis, Lauren Muscott, Linnea Turco, Leyla Tosun, Kelly Atkinson, Elias Assaf, Mini Saraswati, Andy Goodhart, Daniel Kent, Avery White, Will Massengill, Aisha Bradshaw, Margaret Hanson, Marina Duque, Raphael Cunha and Jason Morgan were all an essential part of my graduate student experience, and made the experience immeasurably richer. Special thanks are necessary for Drew Rosenberg and Austin Knuppe: I have had a lot of fun working with you, and I have learned a lot from all of our conversations. Finally, Gregory Smith, my roommate, comps study-buddy, colleague, and friend: I'm so glad I went through graduate school with you. Academic discussions with you have made me a better scholar, but conversations about anything with you have always been rewarding.

My family has supported me throughout this process. I would like to thank my grandfather, Immanuel Wallerstein, for the encouragement during the dissertation-writing process. My uncle, John Lauretig, has been a great source of encouragement. My brother, Joshua Lauretig, has constantly cheered me on. My parents, Mark Lauretig and Susan Morgenstern, have been with me every step of the way and encouraged my academic pursuits; words cannot express how much their support has meant to me. Finally, my beloved wife Anusha has been the best part of graduate school. After meeting for the first time one July day in the Short North Coffeehouse, my life is better with you in it. Not to steal a line, but Johnny Cash put it better than I ever could.

Vita

February 7, 1991	Born in Cleveland, OH, USA
May 2013	B.A., Political Science, Grinnell College
December 2016	M.A. Political Science, The Ohio State University

Publications

Adam M. Lauretig "Identification, Interpretability, and Bayesian Word Embeddings," *Proceedings of the Third Workshop on NLP and Computational Social Science*, (June 2019)

Adam M. Lauretig and Bear F. Braumoeller "Statistics and International Security," *The Oxford Handbook of International Security*, (2018)

Fields of Study

Major Field: Political Science
Political Methodology, International Relations

Table of Contents

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

Since the earliest days of international relations, one of the primary goals of of international relations scholars has been to understand the decisions of policymakers (Morgenthau, 1973, p. 5). However, for quantitative scholars of international relations, elite decision-making has been a black box (Gartner, 1998), they have had to infer the elite decision-making process from observed actions. This means that questions about the relationship between elites and the broader polity, such as whether the elites are simply a conduit for the public's preferences, or whether the public follows elite cues are open questions (see Aldrich et al., 2006 for a review). Elite decision-making, until recently, has remained underexplored by political scientists, who can observe inputs (material circumstances, public opinion) and outputs (policy choices). However, political scientists lack data about how these inputs ultimately shaped the decision-making process, and then, the observed outcome. This is especially problematic for two theories I investigate in this dissertation: casualty sensitivity theory and securitization theory. Elites play a central role in the decision-making process in both of these theories, and yet we know very little about what they do, and the effects of their decisions.

Drawing on the recently digitized *Foreign Relations of the United States*, a collection of declassified diplomatic documents, I test the relationship between elite

decision-making and foreign policy actions using a variety of cutting-edge natural language processing tools. I use a combination of structural topic models (Roberts, Stewart, and Airoldi, 2016) to understand elite casualty sensitivity, and Bayesian Word Embeddings, a model I developed and present in this dissertation (and in Lauretig, 2019), to measure securitization and its policy consequences. In the case of casualty sensitivity, I find that elites escalate bombing in response to casualties. However, they do not change their discussions about the overall direction of the war, suggesting a more complicated picture of elite responses to casualties than the existing literature would have us believe. Investigating securitization, I find that while we see securitization in the cases we would expect, securitizing an issue has no effect on actions, across a variety of outcomes. Overall, these results suggest elite decision-making is more complex than it is initially described in the literature, and this can only be observed by examining the text of foreign policy deliberations. Furthermore, these findings suggest limitations in both quantitative and qualitative approaches to the study of conflict. It suggests the models of conflict used by quantitative scholars distort our understanding of the decision-making process, while the selective use of documents by qualitative scholars can lead us to inflate the effects of any particular decision on policy outcomes.

In this chapter, I review the literature on elite decision-making in American foreign policy. I emphasize the literatures on casualty sensitivity and securitization, and highlight the limitations of current approaches. I then discuss the development of "text as data" methods in political science, and how these methods can help overcome the limitations of existing models of elite decision-making.

1.1 Elite Decision-Making and American Foreign Policy

The relationship between elites and the mass public is central to many debates in understanding American foreign policy, and political behavior more generally. Fundamentally, the debate centers around the question of whether elites follow public opinion, or the public takes cues on issues from elites (Aldrich et al., 2006; Druckman and Jacobs, 2015; Lenz, 2013). The difficulty comes from an inability to measure elite deliberations as they occur. Scholars who want to understand elite decision-making are faced with limited data, especially in crisis situations. There have been survey experiments on elites (Feaver and Gelpi, 2006) and a variety of lab experiments (Hafner-Burton et al., 2014; LeVeck et al., 2014). However, as these experiments may not represent how elites behave in crises, they suffer from concerns about external validity. Furthermore, because elites diverge from the public in a variety of ways, and both diverge from rationalist models (Sheffer et al., 2018), convenience samples from the public may not effectively stand in for elites in crisis simulations.

The limited availability of observational datasets on leaders has meant that many observational studies of leaders focus on institutions, rather than explicitly on leaders, such as the "audience costs" literature (Bueno De Mesquita et al., 1999) which focuses on how variation in institutions affects the propensity of leaders to initiate conflict, and how leaders may be punished by failure on the battlefield (Weeks, 2008, 2012, 2014). Observational studies about leaders have often focused on their backgrounds, particularly, whether they served in the military, and whether this background affects their decisions to initiate conflict (Gelpi and Feaver, 2002; Horowitz and Stam, 2014). What limited quantitative work exists on elite behavior during wartime only characterizes elite behavior with respect to whether a leader is removed from power

or not (Weisiger, 2016). More complicated dynamics are generally not examined, due to a paucity of data.

Elite-centric work that examines the psychological determinants of elite behavior has largely been qualitative, based on archival research. Early versions of this work examined the way cognitive biases shape decision-making in historical case studies of foreign policy, from the role of historical analogies to prospect theory (Jervis, 1976; Levy, 1997). More recent work has extended the psychological examination of leaders, focusing on whether leaders perceive threats as emanating from the institutions of another state, or its actions (Saunders, 2011), how the vividness of threats that leaders perceived affects decision-making (Yarhi-Milo, 2014), and the interplay between presidents and their advisers (Saunders, 2017). A key limitation of this literature is the inability to measure the impact of non-events: qualitative work overwhelmingly focuses on cases where there is documentation of discussion and something happens, it ignores cases where either no discussion occurred, or, where discussion did not lead to action.

Below, I focus on the relationship between the public and elites in the casualty sensitivity and securitization literatures below. I highlight where there are current limitations due to a lack of data on elite decision-making, and what solving these limitations could help us reveal.

1.1.1 Casualty Sensitivity

Initially, scholars believed that the public was apathetic on foreign policy, a view known as the Almond-Lippmann hypothesis (Holsti, 2004). Mueller (1973) showed this was not the case, at least, where casualties were concerned: the public did care

about casualties. Mueller (1973) emphasized that public casualty aversion responded to the logged cumulative sum of casualties, the public cared the most about initial casualties, a finding which has been disputed by Gartner and Segura (1998, 2000), among others, who argue that the public's casualty aversion is driven by more recent casualties, and that more geographically proximate and socioeconomically similar casualties are more relevant to the public's casualty aversion. This aversion mobilizes voters (Davenport, 2015), and thus, we would expect that elites would care about casualties and change their policies, if only to preserve their tenure in office.

However, modeling elite responsiveness to casualties is difficult, especially during conflict, which means that we cannot tell whether public attitudes towards casualties directly shape elite deliberations and behaviors. There is evidence that elites take public backlash into account when planning military operations (Feaver and Gelpi, 2006), however, this research has not been extended to elite deliberations when fighting wars. This has not stopped a mythology from taking hold after the Vietnam War, that the only reason the United States was unable to win was public backlash leading to politicians deciding to end the war (Sorley, 1999). Without access to elite deliberations, we cannot know if this public pressure resulted in the decisions to end the Vietnam war.

An alternate possibility is that elites do not respond to the public; rather, public attitudes are informed by cues from elites (Zaller, 1992). In this literature, changes in the public's views occur when they receive new information from elites. For example, Berinsky (2007) argues that public dissent on war is due to fracturing elite opinion: the public uses elite dissent as their cue about whether the war is going well. However, elite cues are not as passive a process as they initially seem. Presidents in particular

gather a great deal of public opinion data, and choose their cues base on what they think will resonate with the public (Druckman and Jacobs, 2015).

In the casualty sensitivity literature, it is assumed that elites will respond to casualties, whether this response is due to public dissatisfaction or elite dissent. This assumption is due to the absence of easily measured data on elite decision-making, and until recently, an inability to effectively convert archival evidence to data which can test quantitative hypotheses. Thus, scholars have had to infer that casualties lead to decisions to end the war, because wars end as casualties increase. However, they have had to leave elites as a "black box."

1.1.2 Securitization

Securitization theory developed from constructivist theory after the end of the Cold War, in response to the materialist-driven understanding of security promoted by realists at the time. Proponents of securitization sought to understand why some issues become "security issues" and others do not (Wæver, 1995). There was a particular emphasis on issues which were not "classical" security issues, such as the environment (Graeger, 1996) and migration (Huysmans, 2000; Ibrahim, 2005), though, as Buzan (2008) notes, securitization can also be used to understand conventional military issues.

Central to securitization is the idea of the "speech act," which is an utterance composed of what is said (locution), what is meant (illocution), and the effect of an utterance (perlocution) (Austin, 1975). The intent of securitization is to move a particular issue into the realm of security, a realm of "extraordinary politics," where

solving the issue involves claiming any means necessary to solve it (Buzan, Wæver, and De Wilde, 1998; Williams, 2003).

Work on securitization has been almost entirely qualitative (see Caverley and Krupnikov (2017) for the rare exception). This has limited the ability of scholars to measure the effects of securitization on policies; whether the move to securitize an issue results in a measurable change in policy. This is frustrating, as securitization is a "critical cousin" to elite-cue theory, and the ability to measure the effects of securitizing acts could provide a way to compare securitization with mechanisms proposed in conventional elite-cue theories.

The primary limitation to the quantitative study of securitization is the lack of a useful corpus measuring the change in discourse around an issue, and a way to measure the meanings of words. Assembling a corpus of policymaker statements is difficult and time-consuming, especially if not already digitized. However, that is the lesser of two issues. The more difficult issue is measuring the intended meaning, the illocutionary act in speech act theory. Since meaning is conditional and latent, it requires more advanced techniques than standard regression.

1.2 Natural Language Processing and Text as Data

Standard regression-based models fall short in answering questions about elite decision-making. Given that much of the deliberations and policy statements which make up elite decision-making are captured in text, this requires turning to *text as data* tools which have recently emerged in political science.

Text as data can be understood as the application of natural language processing tools to inferential questions, where the goal is to construct variables from text, and

then use these variables to conduct inference (Gentzkow, Kelly, and Taddy, 2017; Grimmer and Stewart, 2013).

Natural language processing involves converting text to numerical representations, such as vectors and matrices, and then, calculating or estimating representations[1] that capture properties of the relationships between words (Turney and Pantel, 2010). The simplest models of this class simply record word frequency, either the frequency with which they appear in a document, or the frequency with which two words co-occur within some distance of each other. The frequency can be weighted using a measure of association like pointwise mutual information, which calculates the empirical probability of two words co-occurring, conditioned on each word's prevalence, or term frequency-inverse document frequency, which weights the frequency of a term by its rarity across a corpus (Church and Hanks, 1990; Manning and Schütze, 1999). However, there is no direct inference from these representations, though they can be used in statistical models.

Inference. However, is possible with probabilistic models, which assume word frequency and co-occurrence is generated from underlying statistical distributions. Topic models are a particularly famous example of this, especially Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003). In Latent Dirichlet Allocation, we assume that each document contains a mixture over "topics," and that words belong to each of these topics with a varying probability. Traditionally, because words are multinomially distributed, these probabilities are drawn from their conjugate prior, a Dirichlet distribution, a distribution of probabilities on the simplex, which means that the probabilities must sum to 1. The output from this model is a set of topics which

[1]Here, I assume calculation is deterministic, while estimation refers to an underlying probabilistic model.

summarize the corpus, and these topics are described by those words which best characterized them. This can be determined by those words which have the highest probability of belonging to a topic, or by a more sophisticated measure, such as the FREX (frequency-exclusivity) score which ranks words according to an average of how unique they are to a topic, and the frequency with which they occur, providing a more coherent characterization of a topic (Airoldi and Bischof, 2016).

While there have been a variety of bespoke topic models for special cases of corpora (Andrews and Vigliocco, 2010; Blei and Lafferty, 2006a,b; Grimmer, 2010; Gruber, Weiss, and Rosen-Zvi, 2007; Mcauliffe and Blei, 2008; Mimno and McCallum, 2008; Wallach, 2006), Roberts, Stewart, and Airoldi (2016) generalize these models by developing the Structural Topic Model, which allows the user to incorporate a variety of metadata about the corpus (such as dates, authors, and other covariates). Furthermore, Roberts, Stewart, and Airoldi (2016) develop a strategy to conduct inference on the topics the model estimates, in order to determine the magnitude of the effect metadata has on topic proportions.

The structural topic model is a powerful tool for conducting statistical inference on text, as it allows users to model the themes in a collection of texts, and how these themes might be affected by other covariates, while providing regression-like interpretable measures of the relationship between metadata and topics.

1.2.1 Word Embeddings

However, topic models make an assumption about the language used in a corpus: that all words are equally (dis-)similar. This is a problematic assumption, especially in the social sciences, where words like “democracy” and “autocracy” are more similar

to each other than words like "democracy" and "kitten." To capture this similarity requires using *word embeddings*. Word embeddings are a set of models which learn which words are most likely to co-occur, and if we assume that words with similiar meanings appear in similar contexts (Harris, 1954), we can leverage this co-occurrence to learn the meanings of words. One advantage of this class of models is that they can capture the similarity of words, even when these words do not directly co-occur, allowing for deeper insights into patterns of meaning in language (Deerwester et al., 1990).

Traditionally, these models would be estimated with matrix factorization. The matrix which was being factorized would use pointwise mutual information (PMI), which calculates the probability of a pair of words co-occurring given their frequency in the corpus. The PMI matrix is then factorized using singular value decomposition (Landauer and Dumais, 1997; Turney and Pantel, 2010).

Word2vec is a modern implementation of word embeddings which provided a fast, scalable embedding model using a single-layer neural network (Mikolov, Yih, and Zweig, 2013; Mikolov et al., 2013). The output from word2vec is a dense numeric vector for each word. However, these embeddings are not directly interpretable. The most common way these embeddings are rendered interpretable is by calculating their similarity using *cosine similarity*, which calculates the cosine of the angle between two vectors, and is bounded $[-1, 1]$. This cosine similarity has been shown to align with human judgement, and also capture biases in language (Garg et al., 2018).

However, the results from word embeddings are simply point estimates, word2vec contains no measure of uncertainty in either the embedding estimates or the cosine similarity measures, nor are they identified. While there have been "Bayesian" word

embedding models (Barkan, 2017; Bražinskas, Havrylov, and Titov, 2017; Ji et al., 2017; Rudolph et al., 2016), they have not taken seriously the role of either inference or identification in estimation.

The ability to develop a probabilistic embedding model, and a way to conduct inference with these embeddings, as demonstrated by Lauretig (2019), is a sizable step forward for combining text as data and theories that rely on understanding word meanings. Bayesian Word Embeddings (Lauretig, 2019) provide a means to link measurements of meaning with standard quantitative tools, and thus, marry interpretivist and quantitative work to measure the effect of securitization and elite discourse on foreign policy.

1.3 Dissertation Outline

In remaining chapters of this dissertation I draw on the *Foreign Relations of the United States* (FRUS), which have been recently digitized, to understand the role of elite decision-making in American foreign policy. I analyze these documents in two ways: first, I use structural topic models to investigate elite casualty sensitivity, and then, I develop Bayesian Word Embeddings to measure the meaning of words, and use this model to investigate securitization.

In Chapter 2, I conduct a detailed examination of elite decision-making in the Vietnam War, the canonical case for casualty sensitivity in American foreign policy (Mueller, 1973). I investigate the effect of casualties on bombing decisions, and find that casualties lead to an increase in bombing. I then examine the effects of casualties on deliberation among elites during the war, using grammatical n-grams and the structural topic model (Handler et al., 2016; Roberts, Stewart, and Airoldi,

2016). I find that casualties have no effect on decisions to end the war, suggesting that while casualties cause elites to retaliate, they do not ultimately bring about the end of a conflict.

In Chapter 3, I then develop a tool to model word meanings more generally: Bayesian Word Embeddings. I address an issue with the suitability of existing embedding models for social science research: embeddings are not identified, and thus, cannot be used for statistical inference. Casting word embeddings as Bayesian latent variable models, I offer a solution: following Rivers (2003), I develop a way to anchor a set of fixed points to identify word embeddings by specifying certain words as endpoints. Additionally, I develop a way to automate this anchoring process, so the user needs to only specify an initial set of points, and then, additional endpoints are specified via the most cosine dissimilar words. I test this model on two corpora: a collection of inaugural addresses, and then, a subset of the FRUS corpus. I find the Bayesian Word Embedding model is able to discover patterns in the inaugural addresses corpus that the Structural Topic Model (Roberts, Stewart, and Airoldi, 2016) is unable to recover, and that a bellicosity measure constructed from the FRUS corpus embeddings have convergent validity with independently coded event data.

In Chapter 4, I apply Bayesian Word Embeddings to understand the role of securitization theory in foreign policy decision-making. After using Bayesian Word Embeddings to construct a measure of securitization, I then test whether securitization has any impact on policy. I investigate three cases: Cuba, after the Cuban Revolution in 1959, France, after it develops a nuclear bomb in 1960, and China, after it develops a nuclear bomb in 1964. I choose these cases because we should expect to see that the securitization of China and Cuba would lead to an increase in conflictual

behavior, but we would not expect to observe this increase with France. However, in all three cases, we see that securitization has no effect on the behavior of the United States with respect to the foreign country, even though the securitizing events affect elite rhetoric. This finding suggests that securitization is a less straightforward process than scholars initially thought, and that it can constrain behavior, as much as it causes new behavior, and this constraining effect is difficult to detect using standard quantitative approaches.

Chapter 2: Elite Responses in Word and Deed to Casualties During the Vietnam War

2.1 Introduction

The shadow of the Vietnam War looms large in American public memory, and in the political science literature. Popular memory holds that college students, upset over the prospect of being drafted to die in an unwinnable war, took to the streets, and their protests over the loss of American life brought the war to an end (Herring, 2008, p. 741). The foreign policy cautiousness after the backlash against the Vietnam war – termed "Vietnam Syndrome" – shaped a generation of policymakers and their foreign policy decision-making (DiCicco and Fordham, 2018). The desire to avoid a backlash from casualties informed a doctrine which favored short decisive wars (DuBrin, 2008), and led to policies to conceal the cost of war from the public, such as banning photography of flag-draped military caskets (Bumiller, 2009). The belief that public outrage from casualties serves as a constraint on elites has potential policy consequences as well, as it promotes regular calls to "bring back the draft" to serve as a check on foreign adventurism (Haberman, 2017; Ricks, 2012).

These calls parallel work in political science which examines the public's sensitivity to casualties during war (Gartner and Segura, 1998; Gelpi, Feaver, and Reifler, 2006; Mueller, 1973). This casualty sensitivity literature assumes that public dissatisfaction

with casualties results in pressure on leaders, who bow to the public, and ultimately, end the war as casualties increase. Many of these studies lean on the Vietnam War as the paradigmatic case for casualty sensitivity influencing elite decision-making (Aldrich et al., 2006).

However, explicit elite decision-making is missing from much of this work because surveying elite attitudes is difficult, and thus rare (but see Feaver and Gelpi, 2006). Existing work can observe casualties as an input, and war termination as output (Weisiger, 2016), but there is no observation of the process by which casualties shape the decision to terminate a war. It is difficult to infer the effect of factors like casualties and public opinion on elites' decisions to terminate a war because this decision-making process cannot be directly measured.

The primary reason for the inability to observe elite decision-making is the lack of detailed data, leading to a mismatch between theory and measurement. Existing studies of changes of public opinion during war have had to either rely on larger time periods for analysis (Gartner and Segura, 1998) or have used fixed events like elections to examine the effects of casualties on turnout (Davenport, 2015). Studies of the differences between individual leaders at war have had to rely on either qualitative case studies (Saunders, 2011; Yarhi-Milo, 2014); large-N studies with coarse measures of leader characteristics have focused on war initiation (Dafoe and Caughey, 2016; Horowitz and Stam, 2014). Using new data, I have the ability to examine how casualties shape elites' daily decisions about both fighting and ending wars, and whether this decision-making process varies across leaders, as recent work has highlighted the importance of individual leaders in decision-making (Saunders, 2011).

In this paper, I provide the first direct investigation of elite casualty sensitivity during war. Using newly digitized documents from the *Foreign Relations of the United States*, daily casualty data, and newly released United States Air Force bombing data from the Vietnam War, I investigate whether casualties cause a change in elite actions and discussions. Specifically, I focus on whether casualties affect decisions about strategic bombing, given the historical evidence showing the hands-on approach of presidents towards bombing, and whether casualties increase the discussion of peace negotiations and whether the war should come to an end.

I find casualties prompted an increase in strategic bombing, but had no effect on the change in discussions around ending the Vietnam War across either administration. Given that these documents include private documents like Henry Kissinger's "backchannel communications," if there is evidence of casualty sensitivity, it should appear in these documents. That casualties do not affect war termination unsettles the existing literature on casualty sensitivity, showing that a key element of the theoretical story – that increases in casualties result in wars coming to an end – does not hold up in the canonical case for casualty sensitivity. While casualties result in both the Johnson and Nixon administrations increasing the intensity of a conflict, ramping up bombing, neither makes any motion towards ending the war.

2.2 Casualty Sensitivity and American Foreign Policy

The belief that elite action is shaped by public disapproval of casualties develops from the long literature on casualty sensitivity among the public. The public disapproves of casualties, and disapproval can be mediated not only by perceived "closeness" to the casualties, but also whether the casualties come from success or

failure on the battlefield. Mueller (1973) was among the first to measure public attitudes around war, and found initial casualties have far more effect on public attitudes than later casualties. However, Mueller's theory did not take into account other events which could inform public opinion. For example, more proximate casualties are more relevant to a given population, with more recent casualties having more of an impact (Gartner and Segura, 1998), and more politically salient casualties–those of the same race, for example–also more likely to shift public attitudes (Gartner and Segura, 2000). Similarly, success matters for the public: publics are willing to tolerate casualties, as long as these casualties are matched by battlefield success (Feaver and Gelpi, 2006; Gelpi, Feaver, and Reifler, 2006). However, the fear of casualties does inspire political activism on the part of the public: parents of men whose sons had lower draft numbers (and were more likely to get drafted), after draft deferments were ended, were more likely to vote in the 1972 election (Davenport, 2015).

The public distaste for casualties has led to a doctrine built around minimizing casualties, through the use of capital-intensive militaries: an embrace of airpower, seapower, and artillery, rather than masses of infantry (Caverley, 2014, ch. 3-4). This doctrine shapes defense spending and procurement decisions (Valentino, Huth, and Croco, 2010), battlefield tactics, and strategic culture. In particular, there is an emphasis on heavy use of airpower on the battlefield (Kocher, Pepinsky, and Kalyvas, 2011), and more generally, a grinding war of attrition, marshaling the industrial and technological resources of the United States to grind down the other side, rather the sacrificing bottomless reserves of soldiers (Weigley, 1977).

2.3 Elites and Casualties

While the public dislikes casualties, they do not ultimately make the decisions to end wars. Existing work supposes that after the public is sufficiently dissatisfied, elites sue for peace, however there is scant empirical evidence for this claim. Much of the research on elite casualty sensitivity focuses on how leaders' backgrounds affect conflict initiation. Gelpi and Feaver (2002) found that American elites who had seen combat were less willing to initiate conflicts than those who had not, and those who had served in the military, but not seen combat are most hawkish (Horowitz and Stam, 2014). When surveyed, military elites (such as flag officers) are again less hawkish than civilian policymakers (Feaver and Gelpi, 2006). However, both groups couch their positions in terms of perceived public casualty sensitivity.

This finding stands in contrast to the elite-cue theories, where public opinion follows elite attitudes (Zaller, 1992). The public will be non-committal on an issue, until an elite division forms, and then, the public will follow the cues of their elite co-partisans. The consensus on the necessity of intervening in World War II and the breakdown of bipartisan support for the Iraq War both follow the predictions of this theory (Berinsky, 2007). The source of the change in elite attitudes is left unsaid. The demonstrated evidence of Vietnam Syndrome, an aversion to initiating conflict, after the end of the Vietnam War suggests again that clearly, there is a mechanism for casualty sensitivity among elites (DiCicco and Fordham, 2018).

In sum, the evidence suggests that American elites are casualty averse when initiating war, and that the effects of casualties linger long after a war is over. However, what is missing from these analyses is actual wartime decision-making from elites: demonstrated evidence that casualties shift either the discussion or the actions of

elites. The assumption of many casualty sensitivity theories, that public outrage over casualties results in elites deciding to end the war, or at least, shift strategy, has yet to be empirically tested.

2.4 Heterogeneity in Elite Responses to Casualties

Much as public casualty sensitivity can be conditional on battlefield success, proximity to casualties, or other factors, elite casualty sensitivity can be mediated by individual characteristics, domestic politics, or structural characteristics of the international system. Heterogeneity in each of these categories could lead to differing responses to casualties. During the Vietnam War, Johnson's Southern background, differing domestic political situations, and the structure of the international system could each mediate Nixon's and Johnson's responses to casualties, with Johnson more likely to pursue a more aggressive bombing strategy, and less willing to negotiate in the face of casualties.

Existing studies of the effect of leaders' background on war have depended on either case studies (Saunders, 2011; Yarhi-Milo, 2014), which limit systematic measurement of the effects of leaders backgrounds, or, have examined the effects of backgrounds on conflict initiation (Horowitz and Stam, 2014), but cannot measure variation in decision-making during war, due to a lack of data. This is a problem for understanding the role of leaders during war because elites have been shown to diverge from rationality in international negotiations, and diverge in different ways that the general public. This divergence from rationality makes modeling elite decisions during wartime difficult without access to elites (Hafner-Burton et al., 2014).

One way elite divergence from rationality can affect wartime behavior is that elites double down, escalating commitments in the face of sunk costs, and elites escalate even more than non-elites (Sheffer et al., 2018). This doubling down can result from elite concerns about reputation (Dafoe, Renshon, and Huth, 2014), and this concern for reputation is particularly relevant for Southern presidents in the United States (Dafoe and Caughey, 2016). Southern presidents are particularly motivated by a "reputation for resolve" which makes them more likely to escalate interstate crises than presidents not from the South, and correlates with an increased propensity to "double down" (Dafoe and Caughey, 2016). The difference between Southern and non-Southern presidents provides one possible explanation for why the Texan Lyndon Johnson might escalate strategic bombings as casualties increased, but the non-Southerner Richard Nixon might not. In his 1966 State of the Union, Johnson illustrated his commitment by vowing to "stay until aggression has stopped" in South Vietnam (Johnson, 1966). Increased strategic bombing provided a means of signaling "resolve," but withdrawing from Vietnam would weaken his reputation for resolve, so casualties would reduce the likelihood that Johnson would enter peace talks.

An alternate theory mediating elite responses to casualties, rooted in the political behavior literature, can be summarized as "elites shape the cues the public takes." While elite-driven theories of public opinion can be traced back to Zaller (1992), these theories generally assume that the public passively follows elite cues. However, Druckman and Jacobs (2015) show that presidents use public opinion polling to choose the issues they talk about, and how they frame these issues. Druckman and Jacobs (2015) highlight how polling on Vietnam shaped presidential decision-making, across Presidents Johnson and Nixon, given the war's prominence in American public opinion

of the time. For Johnson, the goal was to advance his Great Society domestic agenda, helping the poor and creating Medicare. Vietnam was a dangerous distraction from these goals, and he and his administration made several efforts to advance his domestic agenda instead (Druckman and Jacobs, 2015, pp. 100-102). This desire to protect the Great Society could have led to an increase in strategic bombing, as well as less willingness to negotiate by the Johnson administration, as the administration increased strategic bombing in an attempt to staunch casualties, and prevent them from dominating the domestic agenda.

During the Nixon administration, attitudes around Vietnam became increasingly polarized, and voters viewed the Vietnam War as "of the utmost salience." Directly responding to casualties with strategic bombing would undermine the image of a president who put performance above politics that Nixon sought to built up (Druckman and Jacobs, 2015, pp. 54-7, 79-80). Given the preeminence of the Vietnam War in public opinion at the time, retaliatory strikes similar to the Johnson administration's were unlikely to shift public attitudes, and while Nixon attempted to project strength, he also attempted to focus on his foreign policy goals other than the Vietnam War. Thus, while Johnson might turn to strategic bombing as a way to mediate public concerns about the Vietnam War, Nixon faced a domestic political environment where views on the Vietnam had solidified, and were unlikely to change, meaning paradoxically, casualties would not shape his immediate responses as directly, since he was unlikely to change public attitudes about the war.

The structural explanation for the difference between the Johnson administration's and the Nixon administration's responses to casualties draws on classical balancing theory and commitment problems. The Johnson administration faced a bipolar world,

where they believed the Soviet Union was behind all machinations of every Communist government around the world (VanDeMark, 1995). American leaders believed they faced a credible commitment problem: the Soviet and Chinese-allied North Vietnamese represented the advance of Communism in the developing world (Herring, 2008, p. 662), and there was no guarantee it would stop there. In a war which begins from a commitment problem two sides fight because they believe the cost of fighting now is lower than fighting later (Fearon, 1995; Powell, 2006), and keep fighting even as casualties increase. This would lead to an increase in strategic bombing as casualties increase during the Johnson administration, a decrease in willingness to end the war: the increase in strategic bombing missions shows credible commitment in the face of increasing casualties, the United States will keep fighting even as casualties increase.

When Nixon was elected, he faced a different international situation. The Sino-Soviet split occurred in 1969, and Nixon's strategy was less about demonstrating credibility to Moscow, and more about playing the Soviet Union, China, and North Vietnam against each other to secure "peace with honor" (Herring, 2008, ch. 17). Here, showing credibility on the battlefield mattered less than negotiating among multiple actors, suggesting that a straightforward response to casualties would result in Nixon being seen as inflexible and predictable, which could hamper his negotiating position. Thus, while we might expect Nixon to respond less aggressively with bombing to casualties than Johnson, he would be more willing to negotiate, given his stated desire to end the war.

Across all three of these explanations, we would expect that Johnson would be far more aggressive: whether due to bellicosity from his Southern background, the

attempt to shape the narrative in response to battlefield casualties, or differing international constraints, he would be more likely to increase strategic bombing in the face of increasing casualties, and less likely to consider negotiation.

2.5 Casualties and the Vietnam War

The Vietnam War provides the ideal test case for examining casualty sensitivity among elites: there was a significant human cost to the war, there is a public memory of outrage, and there are observable policy changes after the war, due to casualties. Because of the extensive data collection during the war (Daddis, 2011), we can examine multiple responses to casualties: tactical bombing, which are small bombing missions carried out by planes like the A-1 Skyraider; strategic bombing, bombing missions carried out by planes like the B-52 Stratofortress, which bombed larger targets as a part of the broader war effort; and finally, elite deliberations around peace negotiations. These data allow for a detailed analysis of decision-making during the war, examining both patterns of casualties, and the elite responses to these casualties. Below, I explore the arc of the Vietnam War, and the efforts by policymakers throughout the war to minimize casualties.

If there is any evidence of casualty sensitivity in the Vietnam War, we should see it in appear in the historical record, either in the strategic decisions made by elites, or the doctrine and tactics by which the war was fought. Beginning with the initial escalation of the American commitment in Vietnam, and following Nixon's plan to end the war, there were rhetorical moves by both Presidents Johnson and Nixon to empathize with the public's concern about casualties. Furthermore, the war was fought with a set of tactics that emphasized airpower and technology, even when these

tactics proved self-defeating, turning the Vietnamese population against the war (Dell and Querubin, 2017; Kocher, Pepinsky, and Kalyvas, 2011). Airpower involved both tactical bombing missions – air support for soldiers on the ground – and strategic bombing missions, which focus on larger targets, further from the battlefield.

2.5.1 Escalation

After the 1954 Geneva Accords granted Vietnam independence from France, and divided Vietnam into the Communist North and American-backed South, American leaders backed Ngo Dinh Diem, a staunch anti-Communist general (Herring, 2008, p. 662), to build a bulwark against Communism. Through the Eisenhower and Kennedy administrations, this backing took the form of financial backing, military hardware, and a small number of military advisers.

After John F. Kennedy's assassination, Lyndon Johnson continued his predecessor's support for the South Vietnamese government, and escalated the American commitment, as he feared that defeat in Vietnam would imperil his Great Society program domestically, it would "mean and destructive debate that would shatter my Presidency, kill my administration, and damage our democracy" (Herring, 2008, p. 738), in line with the battlefield success argument: the public punishes casualties and *losing* (Feaver and Gelpi, 2006; Gelpi, Feaver, and Reifler, 2006). Prior to the 1964 Presidential election, Johnson refused to escalate, even as Congress (near unanimously) passed the Gulf of Tonkin resolution, which gave him extensive power to prosecute a war against North Vietnam. However, after the 1964 election, as South

Vietnam continued to weaken in the face of the Viet Cong insurgency, Johnson escalated the American commitment, sending 175,000 soldiers in July 1965 (Herring, 2008, p. 739).

2.5.2 Fighting the War

The United States turned to a variety of high-tech solutions to resolve the problems of fighting the Vietnam War. All of these solutions attempted to solve the problem the locating and fighting the Viet Cong and North Vietnamese while minimizing the number of American dead (Karnow, 1994, p. 435).

The policy tool elite decision-makers were most likely to turn to as a substitute for manpower was airpower, in particular, strategic and tactical bombing. Strategic bombing was the use of large bombers like the B-52 Stratofortress to target Viet Cong military infrastructure, and often involved the President and cabinet-level decision-makers choosing targets. Two examples of this presidential involvement are Johnson's "Tuesday Lunches" selecting targets for bombing (Humphrey, 1984), and Nixon pushing for a more punitive bombing campaign against North Vietnam–Operation Linebacker–than the Joint Chiefs of Staff initially wanted (Haun and Jackson, 2016), and turning to aerial mining of Haiphong Harbor to lay siege to North Vietnam. Strategic bombing, then, seems to have been conducted with events on the ground in mind, and with the direct involvement of elite decision-makers, which, if casualties affect elite decision-making, would lead us to believe that casualties would influence strategic bombing decisions during the Vietnam War.

Tactical bombing involved smaller bombing missions to support infantry on the battlefield, where airpower served as rapid-response artillery (Gentile, 2013, p. 72),

suggesting that there would be a relationship between casualties and tactical bombing missions. General William Westmoreland, who commanded the war in Vietnam from 1964 to 1968, was a major proponent of close air support for soldiers on the ground (Malkasian, 2004, p. 936). General Creighton Abrams, who commanded the American military effort 1968-1972, testified to the House Armed Services Committee that tactical air support meant he did not "have to hold a division in reserve," attack aircraft sorties could substitute for a division of soldiers on the ground, and could rapidly be deployed to where fighting was most intense (Abrams, 1973). The emphasis on tactical bombing by military leaders, and the use of airpower as a substitute infantry for suggests that there was an aversion to casualties which shaped tactical decision-making during the war, due to its lower cost in terms of lives.

In both of these cases, the historical evidence leads us to believe that bombing decisions were influenced by casualties at multiple levels of decision-making. In the case of strategic bombing, bombing decisions were often made by the president and his advisers, and if we believe that elite decision-making was influenced by casualties, we would expect to see changes in this strategic decision-making due to casualties. Tactical bombing was a much more rapid response to conditions on the battlefield, and since generals viewed tactical air power as artillery, we should also expect to see a response to casualties in tactical bombing.

2.5.3 Ending the War

Throughout the Vietnam War, there are clear indications that both Presidents Johnson and Nixon were motivated by a belief that public backlash would affect the conduct of the war, and their ability to pursue other policy goals. After the Vietnam

War, the belief that public backlash could constrain adventurism abroad prompted the creation of the selection service, and the requirement that any sustained war required conscription to fight, tying the hands of the president.

Richard Nixon was elected on the promise of "an honorable end to the war in Vietnam" (Nixon, 1968) – later "peace with honor" (Herring, 2008, p. 767). After the shock of the Tet Offensive, Nixon was elected during the most intense period of the Vietnam War, and set about to secure peace with a "secret plan." This plan primarily involved "Vietnamization," transferring the burden of fighting to the South Vietnamese, while providing them with military hardware. The South Vietnamese viewed this plan as the "U.S. Dollar and Vietnamese Blood Sharing Plan." However, it did reduce the American casualty rate (Herring, 1990, pp. 17-18). When Nixon sought to increase military pressure on the North Vietnamese during the Paris Peace Negotiations, he turned to bombing, targeting North Vietnam in Operations Linebacker and Linebacker II, the largest strategic bombing campaigns since World War II (Michel, 1997), and aerial mining, blockading North Vietnam by sea (Herring, 2008, p. 793). While Nixon also invaded Cambodia and Laos, his general approach to ending the Vietnam War involved reducing the number of American soldiers in combat, using airpower to wage the war in place of sending in more soldiers. In his memoir of the war, *No More Vietnams*, Nixon (1985) believed that the public was not willing to pay the cost necessary to win the war, that while the United States lost, it was because of public intolerance for casualties, not that the war itself was unwinnable. Nixon is hardly an objective observer, but the fact that he frames his analysis in the conventional wisdom around casualty aversion shows the deep roots of the belief about public aversion to casualties.

The Vietnam War shaped foreign policy decision-making around war initiation for a generation (DiCicco and Fordham, 2018), as it led to the Powell Doctrine, which favored short decisive wars waged with overwhelming force for a clear goal (DuBrin, 2008). Military elites, after the end of the Vietnam War, were willing to translate their casualty aversion into policy guidance (DiCicco and Fordham, 2018).

2.6 Hypotheses

Theories of casualty sensitivity, elite behavior, and the historical record offer five hypotheses to examine the effect of casualty sensitivity on elite decision-making. The first hypothesis is about tactical behavior, examining policy choices which were enshrined in doctrine, and the second and third more directly examine intentional elite behavior: how casualties shape the strategic decision-making of elites during war. The fourth and fifth are mediating hypotheses, asking whether Johnson would respond differently to casualties than Nixon.

The first hypothesis is about rapid response to casualties: whether casualty sensitivity is enshrined at the doctrinal level. Generals believed that airpower was essential for battlefield success (Malkasian, 2004), and that airpower could substitute for divisions of infantry. If casualty sensitivity affects immediate tactical decision-making we would expect that:

H1: *An increase in casualties results in an increase in tactical bombing missions.*

To investigate the effects of casualties on elite decision-making, we can investigate what elites do, and what they say. To investigate what elites do, we can examine strategic bombing decisions, those missions carried out by larger bombers targeting infrastructure, rather than battlefield targets. These missions were often directed

by the President, with both Lyndon Johnson and Richard Nixon picking targets, and overruling the military's choice of targets (Haun and Jackson, 2016; Humphrey, 1984). Thus, strategic bombing is a response directly controlled by elites in the face of casualties. If elite decisions are affected by casualties, we would expect to see an increase in strategic bombing with an increase in casualties:

H2: *An increase in casualties results in an increase in strategic bombing missions.*

While an increase in casualties may lead to an increase in bombing, this tells us little about whether the war is more likely to end. To do that, we would want to know how casualties shape decisions to end the war. Since we cannot poll elites about their attitudes during wartime, observing their internal discussions allows us to gain some leverage on whether casualties affect elite attitudes about the course of the war more broadly, beyond bombing decisions. If standard casualty sensitivity theory applies, we would expect that an increase in casualties would increase discussions around ending the war, because elites believe the public is less willing to stomach the costs of war:

H3: *An increase in casualties results in an increased discussion of ending the war.*

The fourth and fifth hypotheses are mediating hypotheses, drawing on the literature on leader psychology and domestic politics. Due to Johnson's Southern background, the demands of domestic politics, and the differing international situation, we might expect that Johnson would react more aggressively to casualties, and be more willing to escalate, and less willing to negotiate than Nixon:

H4: *In the Johnson administration, an increase in casualties results in a larger increase in strategic bombing missions compared to Nixon administration.*

H5: *In the Johnson administration, an increase in casualties results in a decreased discussion of ending the war compared to the Nixon administration.*

2.7 Data

Existing approaches to modeling casualties and wartime decision-making during war face several limitations. War-level datasets, like the Correlates of War (Sarkees and Wayman, 2010), which only report onset, duration, outcome, and magnitude, make it impossible to model the decision-making processes during war. While datasets of battles exist (Biddle and Long, 2004), and have been used to model battlefield decision-making, these datasets suffer from selection bias (over-reporting wars like World War Two), and inconsistent coding of battles. Weisiger (2016) provides the most comprehensive coverage of battlefield deaths during war. However, there is no other measure of conflict intensity, nor any direct link to strategic decision-making.

Focusing on the Vietnam War allows for detailed examination of casualties, strategic and tactical responses to casualties, and policymakers' deliberations in the face of casualties. This is possible because the Vietnam War was thoroughly documented (Daddis, 2011). Not only are there detailed daily casualty records, there is also detailed bombing data, differentiated by mission type, which allows another measure of conflict intensity. These data allow for detailed, daily-level analysis of elite responses to casualties. Existing studies of changes of public opinion during war have had to either rely on larger time periods for analysis (Gartner and Segura, 1998) or use fixed events like elections to examine the effects of casualties on turnout (Davenport, 2015). Instead, I have the ability to examine daily-level discussion among elites, and the effects of casualties on discussions, allowing for "real-time" analysis of war-time decision-making.

2.7.1 Casualties

Casualties, the key independent variable, are taken from the Department of Defense (*Defense Casualty Analysis System* 2008). Deaths are summed by day, and aggregated across services. Figure 2.1 displays a clear increase in deaths during the Johnson administration, and a decrease during the Nixon administration. The Johnson administration averaged 20 casualties per day (sd = 23.6), and the Nixon administration, 14 (sd = 16.3). On the worst day for the Johnson administration, the United States suffered 246 casualties, on the worst day of the Nixon administration, 209. Trends in bombing and casualties look similar to the naked eye, particularly during the Johnson administration, as both bombing and casualties increase through the first half of 1968, and decrease afterwards.

2.7.2 Bombing Missions

To measure policy decision-making during the Vietnam War in response to casualties, I use detailed data on air missions during the Vietnam War. These data are available from the Air Force through the Theater History of Operations Reports (THOR) database, allowing a fine-grained measure of conflict intensity (Robertson, 2013). Missions are subsetted to those with clearly offensive intent, cargo flights, reconnaissance, escort, and refueling flights are removed. The remaining missions are then subsetted to missions with a recorded release of weapons. These are selected using regular expressions, which, since Air Force aircraft follow a relatively uniform naming pattern (fighters are "F-", bombers are "B-", attack aircraft are "A-"), provides a consistent pattern to extract. I categorize tactical bombing missions as those conducted by attack aircraft, such as the A-1 Skyraider, and strategic bombing

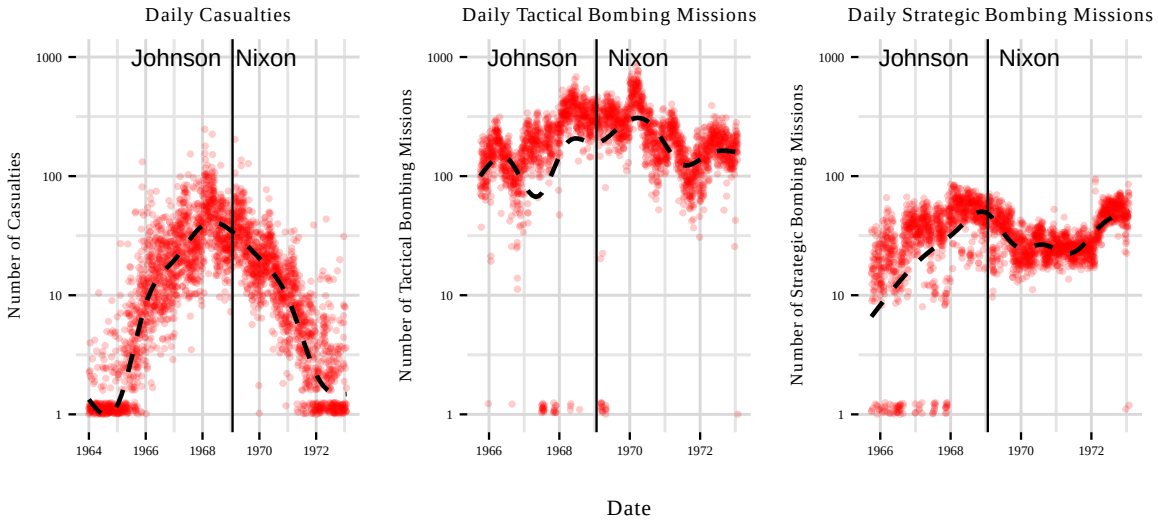

Figure 2.1: Daily casualties and counts of tactical and strategic bombing missions. The vertical line represents the change between Johnson and Nixon administrations. Dashed lines indicate loess-smoothed time trends.

missions as those conducted by bombers, such as the B-52 Stratofortress. Missions are aggregated by day, and the daily sum of bombing missions is plotted in Figure 2.1. On average, Johnson oversaw 204 (sd = 117.5) tactical bombing missions and 34 strategic bombing missions per day (sd = 18.3), and Nixon oversaw 223 tactical bombing missions (sd = 117.6), and 33 strategic bombing missions (sd = 12.4).

2.7.3 The *Foreign Relations of the United States* Corpus

To analyze the topics of elite discussion, I rely on the documents collected in the *Foreign Relations of the United States*. The *Foreign Relations of the United States* (FRUS) contains primary source documents of private communications from the policymakers who develop and implement the United States' foreign policy. Among the sources for documents included in FRUS are "Presidential libraries, Departments of State and Defense, National Security Council, Central Intelligence Agency, Agency

for International Development, and other foreign affairs agencies as well as the private papers of individuals involved in formulating U.S. foreign policy," with a focus on documents relevant to policy-making (*About the Foreign Relations of the United States Series*). When a FRUS volume is compiled, the compiler(s) first identify a set of themes, develop a list all relevant documents, and then select those with the greatest historical import. These are then redacted or declassified, typeset, compared to the original document, printed, and bound (McAllister et al., 2015).

This corpus, then, collects the most politically relevant documents surrounding a particular region or subject, arranged according to country or theme by year. The hand selection by State Department historians has removed the majority of documents (numbering in the millions), which, as Chaney et al. (2016) note, are routine minutiae, such as requests to visit American towns. Inferentially, this is helpful to political scientists, they can focus on the documents which shape policy. One downside to relying on the FRUS dataset, however, is that we lack a baseline: the ratio of "policy relevant" cables to "irrelevant" cables; additionally, we lose the ability to determine the prominence of particular actors. The only way to determine this would be to gain access to the universe of FRUS cables, which is not available for the entirety of the Vietnam War. If casualties do shift the topics of elite discussions, this corpus would likely cause us to over-estimate the effect of casualties on the topics of elite discussions, since it is likely that casualties would cause changes in policy discussions, but not boilerplate discussion about embassy logistics and passports, for example.

I focus on documents from 1964 to 1973 in the Vietnam War FRUS collections, from the beginning of the year that marked the start of major combat operations in Vietnam to the signing of the Paris Peace Accords on January 27, 1973, ending

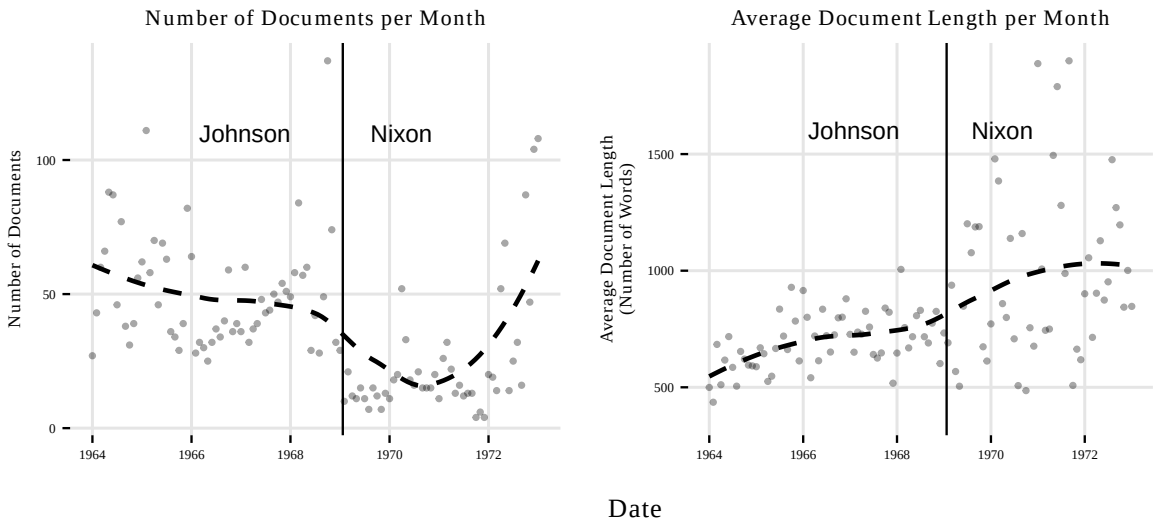

Figure 2.2: Monthly counts of documents, and average length.

direct American involvement in Vietnam. These documents are particularly useful for scholars interested in elite decision-making: they represent an unfiltered view into private discussions among elites in "real time." Documents were extracted from the digital FRUS volumes, and tokenized: each document was turned into a list of individual words, with stray letters, punctuation, and words associated with the archival nature of the corpus removed. While a complete set of descriptive statistics for the corpus are discussed in the appendix, Figure 4.1 shows, I show that the average number of documents per month ranges from 25-75, and that during the Nixon administration, while there are fewer documents included, the average length of each document increases.

An initial examination of the contents of the FRUS Vietnam corpus suggests that these documents capture debate of interest to scholars, in documents which reveal policymakers' private attitudes towards the war and negotiations. An initial

examination of words related to combat ("death," "casualties," "casualty," "battle," "fight," "insurgent," "bomb," "attack") and peace negotiations ("peace," "treaty," "negotiate," "negotiation," "treaties") in Figure 2.3 shows that more than half of the documents per month discussed the war, which highlights how central the war was to elite discussions. Examining the frequency of negotiation-related words, we see increases in 1968-9, and 1972-3, which accord with the historical record: the Johnson and Nixon administration's efforts to negotiate an end to the Vietnam War.

Furthermore, the types of documents contained in the corpus suggest that they are representative of private elite discussions. There are 61 types of documents in the FRUS Vietnam corpus, determined by the first two words of the document metadata, since this metadata appears in a consistent format, such as `Message from the Ambassador in Vietnam ( Lodge ) to the President`. In Figure A.3, the ten most frequently occurring document types are shown, and it is clear that internal, intra-elite communication comprises the majority of the documents in the corpus, and, based on their types, they represent private communication. Approximately 60% of the documents are memoranda, telegrams, or messages, with senders and receivers, addressing the creation and implementation of policy. A smaller number of the documents are summaries, reports, diary entries, or action memoranda, without a clear direction.

The "National Security" category refers to National Security Action Memoranda, National Security Studies, National Security Decisions, documents produced by the National Security Council. The "Action Memorandum" category is also produced by members of the national security council, and contains sender-receiver information in

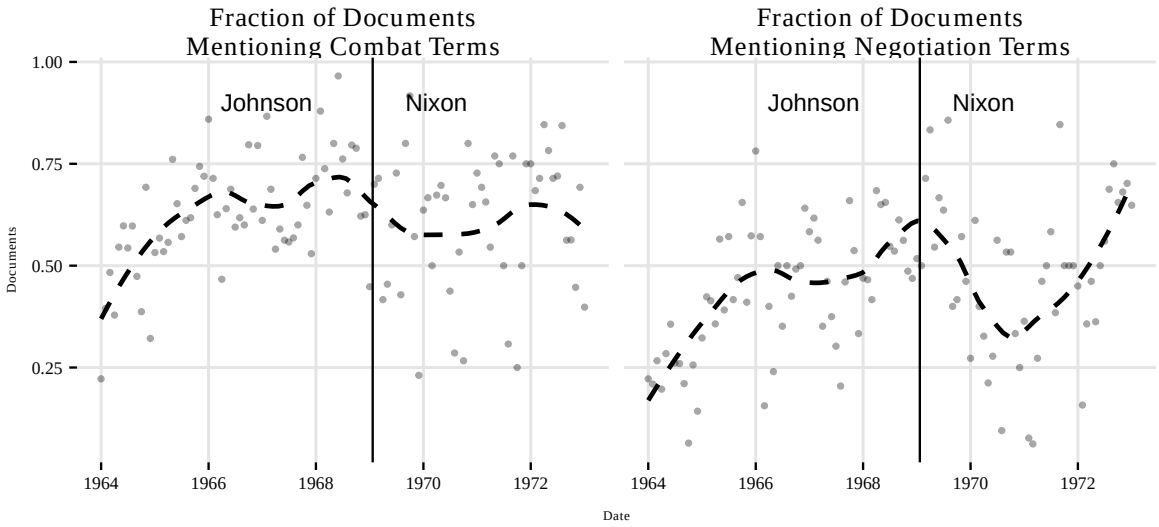

Figure 2.3: Percentage of documents per month which discuss battle related words or peace related words.

the metadata. The "Memorandum Prepared" category are mostly prepared by the Central Intelligence Agency, and do not contain directional information.

One category, "backchannel messages," stands out and helps confirm that these documents do capture the private interactions of elites which are most likely to reveal their attitudes. These documents were communications between the White House (Nixon and Kissinger) and U.S. ambassadors, but kept secret from the official State Department bureaucracy. Similar backchannel messages were used during the Paris peace negotiations, and provide a more accurate view of policy decisions than "official" cables (*Foreign Relations of the United States Guide to Sources on Vietnam, 1969-1975*). The inclusion of these messages helps to assuage concerns that classification would remove all of the truly private communications between elites, and these backchannel messages include discussions of bombing strategy, training Vietnamese allies, and peace negotiations.

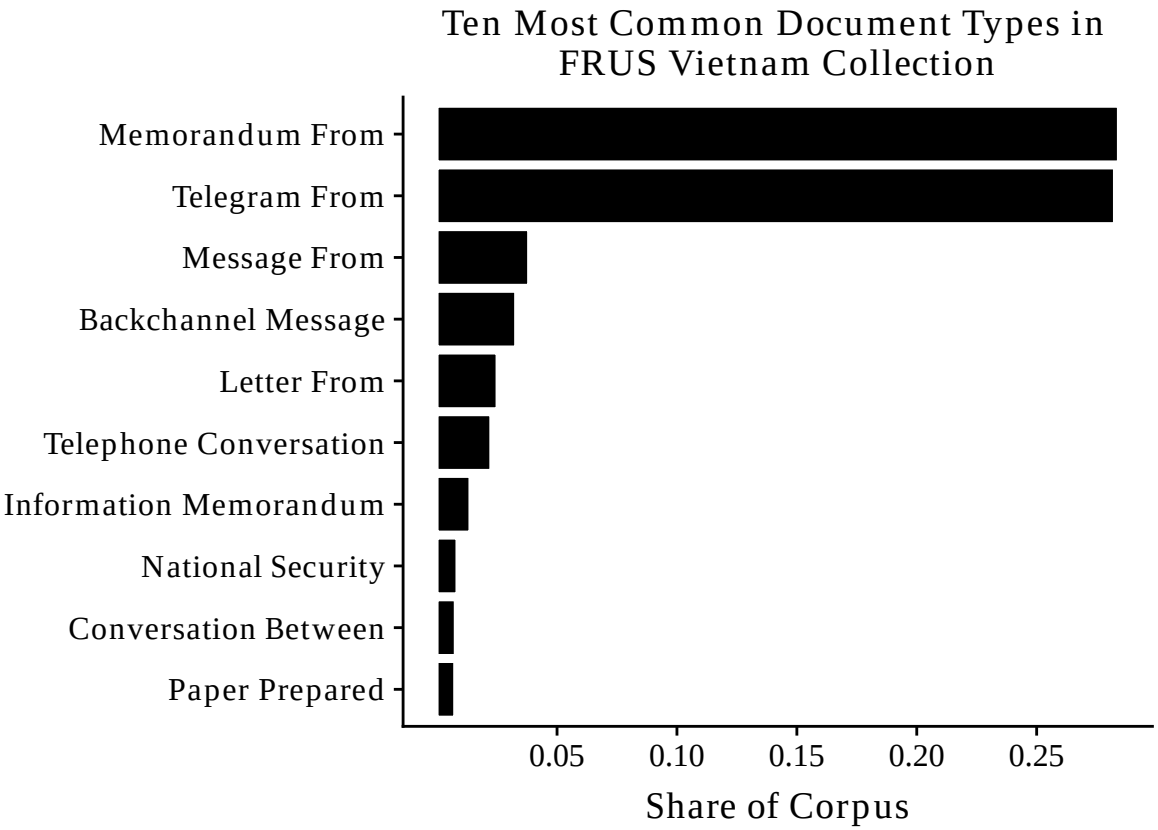

Figure 2.4: 20 Most Frequent Document Types in the Foreign Relations of the United States Corpus.

2.8 Do Casualties Affect Bombing Missions?

I begin by investigating actions in response to casualties, namely the response in bombing to an increase in casualties, testing **H1** and **H2**. To operationalize these hypotheses, I turn to the casualty and bombing data discussed above.

To test the effect of casualties on bombing decisions, I first regress tactical bombing missions on the casualties from that day and the previous day's tactical bombing missions. I then regress strategic missions on casualties from the previous day. To account for variation in the intensity of the Vietnam War over time, I include yearly fixed effects, and to account for possible differences in presidential strategy, I include an indicator for Johnson's presidency vs. Nixon's, and interact this indicator with the lagged casualty variable in the strategic bombing equation. I also include the previous day's tactical bombing in an additional model, as a second measure of conflict intensity, and display the results in Table 2.1. I also include two models with variables capturing domestic (American) economic trends, as captured in inflation and unemployment.

We see evidence supporting **H1**, tactical bombing missions increase with casualties, every eight casualties results in an additional tactical bombing mission. This accords with the historical evidence above, that there was a heavy reliance on airpower to support troops on the ground, and if eight deaths (slightly smaller than a standard squad of soldiers) could result in an additional airstrike, that commanders were quick to turn to airpower in the heat of battle. What is surprising is not that battlefield decision-makers are casualty sensitive, but rather, the degree to which they are sensitive. When there are more than two hundred tactical bombing missions per day, the decision to turn to capital-intensive happens relatively quickly. This helps

	Tactical Bombing	Strategic Bombing	Strategic Bombing	Strategic Bombing	Strategic Bombing	Strategic Bombing
Casualties	0.12*					
	(0.06)					
Previous Day Casualties		0.06*	0.05*	0.05*	0.06*	0.05*
		(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Previous Day Attack Missions	0.84*		0.03*	0.03*		0.04*
	(0.01)		(0.00)	(0.00)		(0.00)
Nixon	−12.36	−14.49*	−12.42*	−12.14*	−15.08*	−13.59*
	(11.35)	(2.60)	(2.52)	(2.69)	(2.79)	(2.70)
Previous Day Casualties × Nixon				−0.01	0.00	0.01
				(0.03)	(0.03)	(0.03)
Inflation Rate					0.98*	1.27*
					(0.42)	(0.41)
Unemployment Rate					−1.95*	3.06*
					(0.85)	(0.89)
Intercept	19.02*	13.28*	9.92*	9.89*	19.92*	−4.88
	(5.22)	(1.16)	(1.16)	(1.17)	(3.69)	(3.99)
R^2	0.83	0.48	0.51	0.51	0.48	0.52
Adj. R^2	0.83	0.48	0.51	0.51	0.48	0.52
Num. obs.	2675	2676	2675	2675	2676	2675
RMSE	48.06	11.05	10.69	10.69	11.04	10.65

*$p < 0.05$. Year fixed effects are suppressed for clarity.

Table 2.1: Effect of Casualties on Number of Strategic and Tactical Bombing Missions

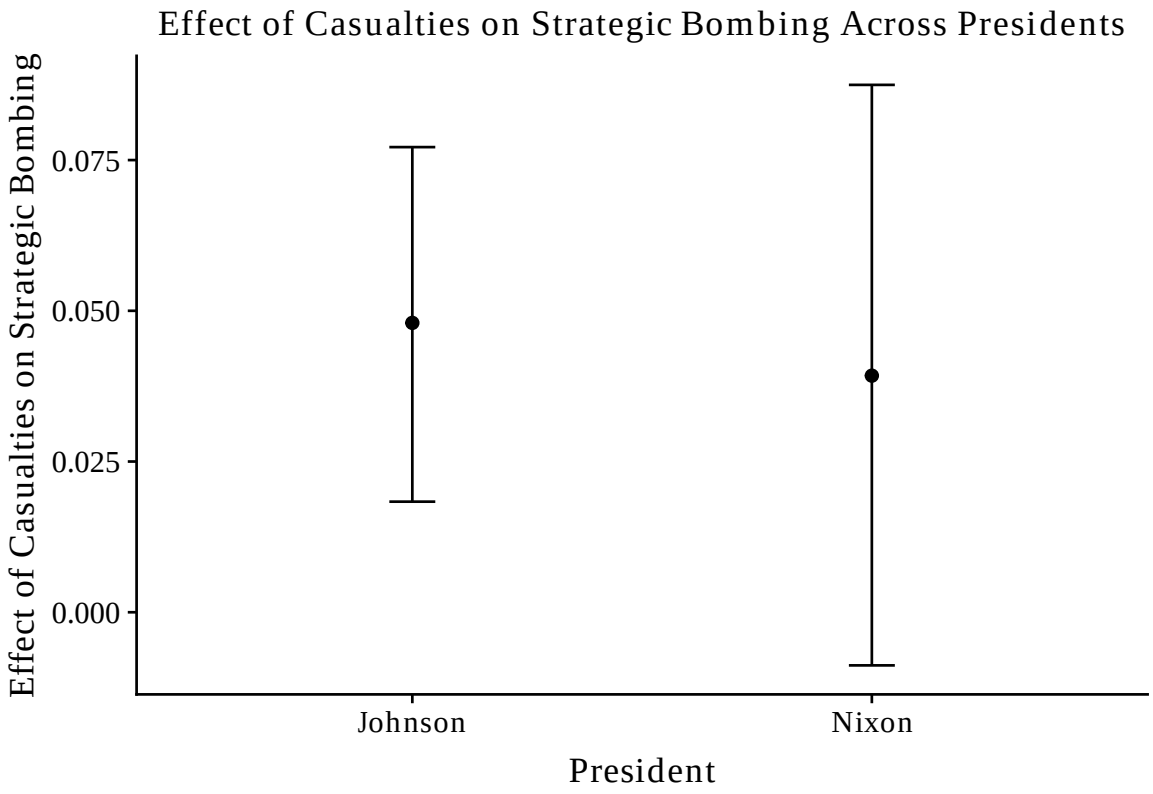

Figure 2.5: Average marginal effect of casualties on strategic bombing across presidents.

confirm that casualty sensitivity played a role in battlefield decision-making, and that, on the battlefield, commanding officers were quick to turn to airpower in the face of casualties.

Examining the strategic bombing models testing **H2**, we again find support for the hypothesis, with an increase of approximately 20 casualties in the previous day resulting in an additional strategic bombing mission. I also test whether there are differences in casualty sensitivity between the Johnson and Nixon administrations and display the results in Figure 2.5. The difference between presidents is not significant (Gelman and Stern, 2006), and thus, we fail to reject the null hypothesis for **H4**.

With an average of 33 strategic bombing missions per day, this result suggests that elites react more strongly to an increase in casualties than commanders on the ground, since a one standard deviation increase in casualties during the Johnson administration (an increase of 23.6 casualties) results in an additional bombing mission, a 3% increase, while the same increase in casualties results in only an additional 2.83 tactical bombing missions, which is only a 1% increase in tactical bombing missions.

We can further examine the patterns of casualty sensitivity in the Johnson administration's war effort because we know the scheduling of the decision-making process. Bombing targets were picked during the "Tuesday lunches," which means that the effect of casualties on bombing decisions can be observed through these meetings. I do this here, summing casualties and bombing missions by week, with weeks running from Tuesday to Monday, so that each new week begins on a Tuesday. I then regress the strategic bombing totals for each week on the previous week's casualties. I also include a weekly time trend, year fixed effects, and the summed tactical bombing missions from the previous week, and display results in Table 2.2.

	Strategic Bombing	Strategic Bombing	Strategic Bombing	Strategic Bombing
Previous Week Casualties	0.10*	0.12*	0.05*	0.09*
	(0.02)	(0.02)	(0.02)	(0.02)
Previous Week Attack Missions			0.07*	0.07*
			(0.00)	(0.00)
Weekly Trend		0.37*		0.83*
		(0.14)		(0.13)
Intercept	90.37*	86.90*	37.05*	24.89*
	(7.51)	(7.61)	(7.50)	(7.61)
R^2	0.63	0.63	0.70	0.71
Adj. R^2	0.63	0.63	0.70	0.71
Num. obs.	1203	1203	1202	1202
RMSE	68.24	68.08	61.43	60.39

*$p < 0.05$. Year fixed effects are suppressed for clarity.

Table 2.2: Effect of Casualties on Bombing Missions (Tuesday Meetings)

The effect of casualties on bombing follows the same pattern observed in Table 2.1, and the pattern holds across the various model specifications of the Tuesday meetings models. The model can be understood as an additional 20 casualties in the previous week resulting in an additional 1-2 strategic bombing missions in the present week, with this effect remaining stable across all specifications. This result helps reinforce the idea that elites were responding to casualties in their decisions in increase strategic bombing, and that casualties prompted elite escalation.

To further understand the effect of casualties on the Johnson administration's decision-making, I plot the effect of casualties on predicted strategic bombing decisions holding all other variables constant, and use a smoother to predict bombing missions from the entire range of casualties observed at the weekly level during the Johnson administration in Figure 2.6. In Figure 2.6 we see that increasing casualties result in an increase in strategic bombing missions, and that, in particular, the initial increase in casualties, and a large number of casualties produce the largest increases in the number of strategic bombing missions.

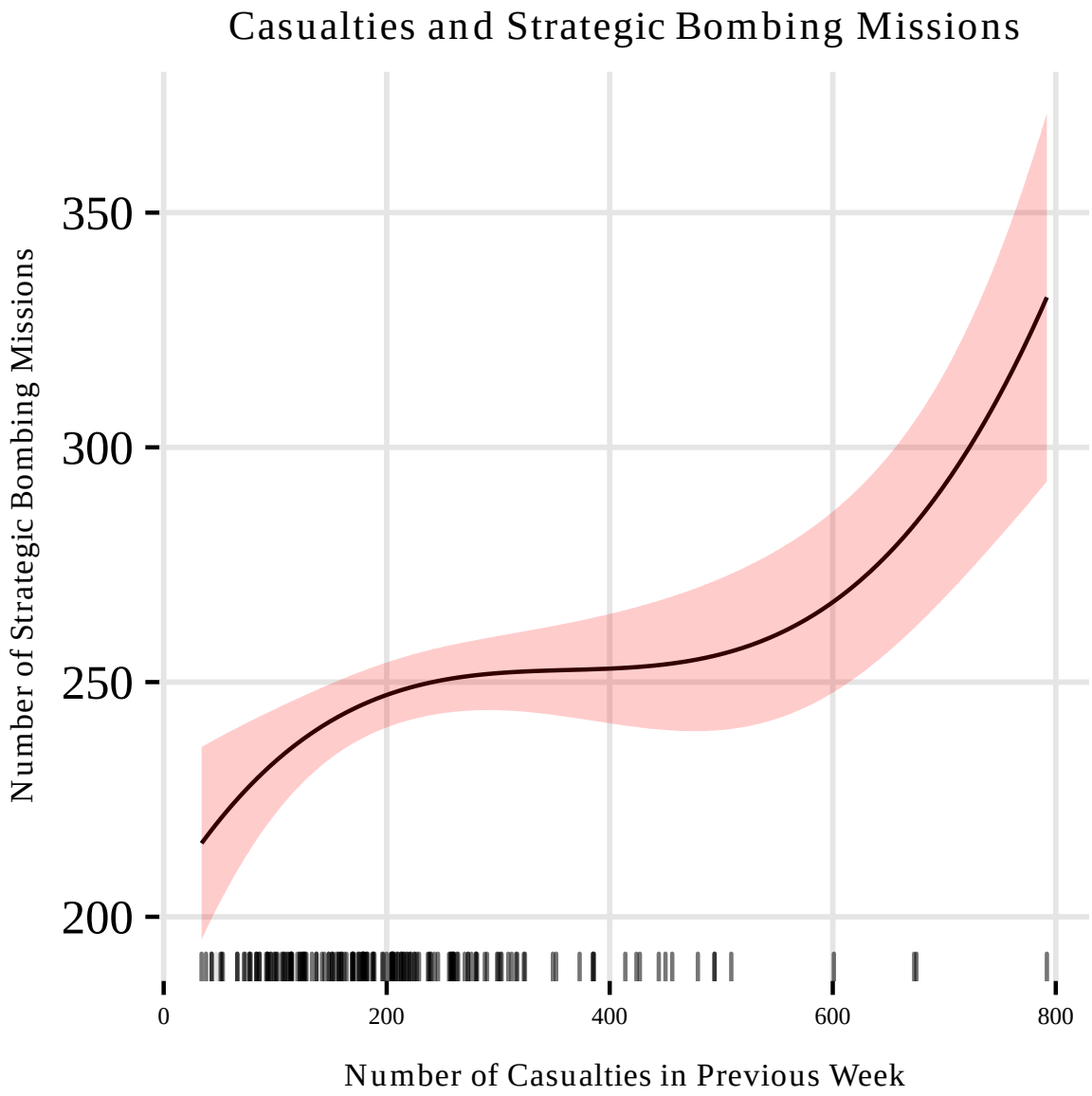

Figure 2.6: Estimated strategic bombing response to the previous week's casualties from Johnson administration's Tuesday meetings process.

2.9 Do Casualties Affect Elite Discussion?

While examining the effect of casualties on bombing decisions can illuminate the decision-making process during a war, it cannot explain why a war comes to an end. To understand why wars end, we must examine how casualties affect the decision-making process by elites: any war that does not end with the annihilation of one side ends because of a choice by decision makers. One way of measuring whether casualties affect decisions to end to war is measuring whether casualties cause the elites to increase discussions of ending the war. The FRUS corpus allows us to observe what elites discuss as events happen: if an increase in casualties correlates with an increase in discussions about peace negotiations and ending the Vietnam War, we can take this as evidence of casualties influencing the decision-making process.

The FRUS corpus provides a record of elite discussion in real time, and a potential source for a dependent variable to analyze these elite discussions: the topics of elite discussions. I focus on topics, rather than individual words, because topics can capture words with similar meanings that are shared across documents.

To measure the content of elite discussion, I focus on topic models (Blei, Ng, and Jordan, 2003), which estimate latent themes in a corpus of documents based on the frequency of word co-occurrence across documents. Words which co-occur more often across documents are used to estimate topics, underlying latent themes in the corpus of documents, and words belong to topics with varying probabilities. For example, "insurgency," "pacification,", "counter-revolutionary", and "rural development" might all be captured in a topic about fighting the Viet Cong. Documents are represented by their topic proportions, and words belong to some topics with higher probabilities than other topics. By examining the words most associated with each

topic, we can then gain an understanding of what elites discussed during the Vietnam War as the war happened.

Measuring the effect of casualties on elite discussion necessitates using the structural topic model (STM), a descendant of the topic model which allows the user to incorporate covariates into topic estimation (Roberts, Stewart, and Airoldi, 2016). The STM also provides a way to estimate the effect of covariates on topic proportions, in this case, the effect of casualties on the topic proportions in the FRUS corpus, analogous to a regression of topic proportions on a covariate of interest. If the substance of discussions changes due to casualties, this would be reflected in the output from the structural topic model: as casualties increase, there would be more change in the topic proportions under discussion.

However, there are limitations to topic modeling for understanding elite discourse. Topic models make a "bag of words" assumption about the documents: words are single tokens and are assumed to be exchangeable, all that matters is their frequency, not the order in which they appear (Blei, Ng, and Jordan, 2003). I leverage grammar to construct meaningful, multiword noun-phrases, an approach developed by Handler et al. (2016). An example from the corpus under study: rather than counting "north," "south," and "vietnam" independently, combining the adjectives with Vietnam into "north vietnam" and "south vietnam" is far more informative, these combinations are referring to different countries, with opposite meanings.

2.9.1 Structural Topic Model Specification

I estimate the topic proportions using a model similar to the models specified in Table 2.1, allowing comparison between the effects of casualties on actions and

words. I include several covariates in my model: casualties from the previous day, tactical bombing missions from the previous day, strategic bombing missions from the previous day, document date, and under which presidential term the document was written. I interact the previous day's casualty counts with an indicator variable for the president at the time. As Figure 2.1 shows, the trends in casualties are almost exactly opposite across administrations, and not accounting for this could damage inferences. In the supplementary materials, I also test the effect of casualties and missions from the previous week, and find the same effect.

Additionally, because there might be a difference between those documents which involve elected officials and those which involve career officials (public servants, military officers, etc.), I subset the corpus to all those documents which have the term "president" in their metadata (sender, receiver, document type), as they may be fundamentally different from those in the broader corpus. I run the model on the full corpus in the supplementary materials, and the results are substantively the same.

To recover more coherent topics, I fit a model with a large number of topics (in this case, 100). With a large number of topics, a few (5-10) topics will absorb high probability words which contribute little to identifying topic meaning, and exhibit stop-word like behavior, leaving the rest of the topics as substantive topics for interpretation (Wallach, Mimno, and McCallum, 2009).

2.9.2 Estimated Effects

I examine whether casualties increase the proportion of elite discussion about various topics surrounding the end of the war. The words and phrases most associated with each topic are displayed in Table 2.3, ranked by their frequency-exclusivity

(FREX) score: those words and phrases which are both most unique to a topic, and which occur most frequently in a topic; this score can be thought of as those words and phrases which best characterize a topic. There is a general topic about peace negotiations, and topics capturing both the Johnson administration's 1968 peace negotiations, and the Nixon administration's negotiations around the Paris peace accords.

To test **H3**, I regress the proportions for the topics in Table 2.3 on lagged casualties, interacted with the Presidential indicator, holding all other variables constant. In Figure 2.7, I report the results from regressing the change in the peace negotiations topic on the previous day's casualties, across both the Johnson and Nixon administrations, and find that casualties have no effect on the change in proportion of discussion captured by the peace negotiations topic.

I then investigate the effects of casualties on administration-specific negotiation topics, and present the results in Figure 2.7, and again find that casualties have no effect on the change in the topics of elite discussion. What stands out is how precisely estimated the null effects are. When the number of casualties is greater than zero, the change in topic proportions statistically indistinguishable from zero, and the 90% confidence intervals also contain zero, further strengthening the idea that there is no change in the topics of discussion from an increase in the previous day's casualties (Rainey, 2014). Furthermore, I fail to reject the null hypothesis for **H5**, the reponses by both the Johnson and Nixon administrations to casualties are not significantly different across the shared peace topic and the administration-specific topics.

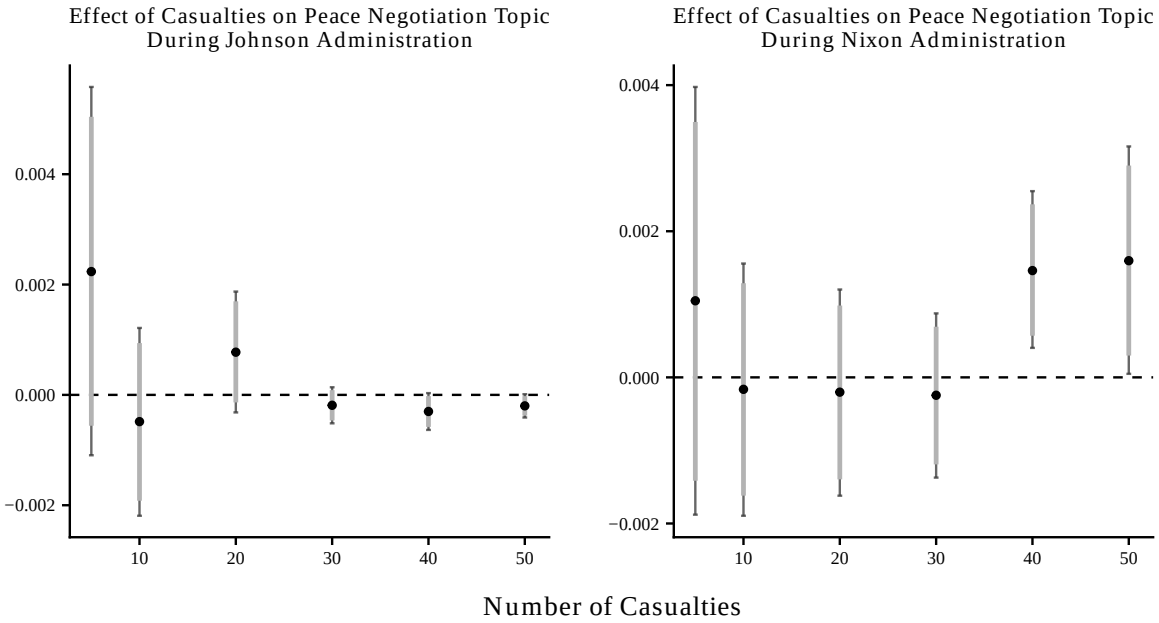

Figure 2.7: Effect of an increase in casualties on the change in the Peace topic across the Johnson and Nixon administrations. Plot shows both 90 and 95% confidence intervals.

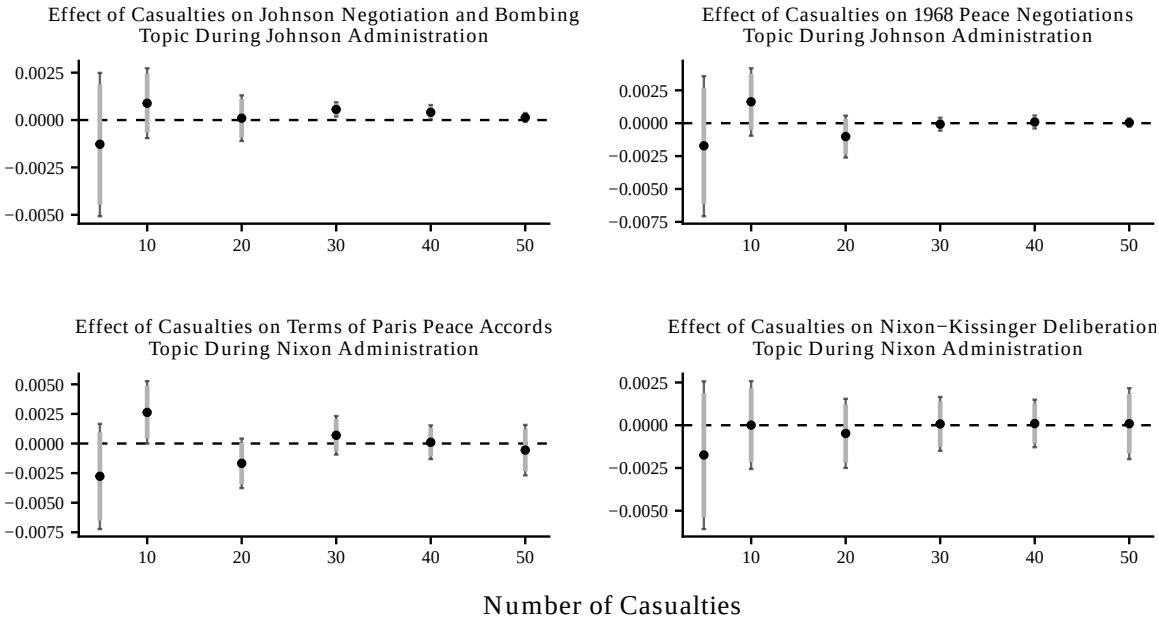

Figure 2.8: Effect of an increase in casualties on the change in several negotiation topics across the Johnson and Nixon administrations. Plot shows both 90 and 95% confidence intervals.

Peace Negotiations	Johnson Negotiation and Bombing	1968 Peace Negotiations	Terms of Paris Peace Accords	Nixon-Kissinger Deliberations
territorial integrity	reconnaissance	san antonio formula	general haig	tohak whs
independence sovereignty	deployment	mcnamara ball	negotiators	saturday sunday
military alliance	gia lam	antonio formula	points entry	same spirit
rainy season	units	san antonio	national council	posture
internal affairs	decision	hard core	sufficient time	tomorrow afternoon
access	air offensive	many ways	one hour	open break
communist government	acceptable settlement	living conditions	administrative structure	side honor
such use	rolling thunder	important matter	council national reconciliation	settlement package
top secret	phuc yen	such contacts	delegation	such extension
new government	attacks	use .s	private sessions	local time
one thing	increase	hanoi radio	national council national reconciliation	one week
television interview	policy	mainland china	saturday morning	logistics system
honorable settlement	divisions	such plan	northern part	technical experts
two zones	air squans	political sense	xuan thuy	kissinger hanoi
maximum goodwill	line pause	negotiating table	willingness	spirit good will

Table 2.3: Selected topics, top 15 FREX-scored phrases from the topic model run on daily-level data with a previous-day casualty lag, on documents associated with the President, with human-coded labels.

2.10 Conclusion

In this paper I set out to understand how casualties affect the decision-making of elites during the Vietnam War, and to expose a previously-missing link in the theoretical model of casualty sensitivity. Using daily-level casualty, tactical bombing, and strategic bombing data, I found that casualties provoked a statistically significant increase in tactical bombing across both presidents during the Vietnam War. Taking advantage of the known schedule of the Johnson administration's decision-making process, I was able to further confirm the effects of casualties on elite bombing decisions. Surprisingly, I also found that casualties did not change the amount of discussion about ending the war, in contrast to what casualty sensitivity theory would have us expect.

This result problematizes a previously untested assumption in the current theoretical model of casualty sensitivity: casualties, and the resulting public outrage, bring about the end of a war. While a variety of works have focused on the role of public opinion, and used the Vietnam War as their case (Gartner and Segura,

1998; Mueller, 1973), the assumption that leaders respond to this outrage has largely remained untested. Feaver and Gelpi (2006) have shown in experiments that elite perceptions of the public's casualty sensitivity matters for decisions to initiate wars. However, I have shown that this logic does not apply to decision-making to end wars. Furthermore, by examining strategic bombing decisions in response to casualties, I show that some elites do respond to an increase casualties by further escalating the conflict. This result suggests that while some elites are sensitive to casualties, it only affects decisions to retaliate and escalate, not larger decisions to end the war.

More generally, this finding unsettles much of the literature on conflict termination, which assumes that casualties are the mechanism by which wars come to an end (Weisiger, 2016). While there is certainly correlational evidence that an increase in casualties results in the end of a war, this is not borne out by observation of leaders' decision-making. If casualties are not the direct mechanism by which wars come to an end, we must rethink our understanding of the relationship between domestic politics, elites and war termination. We must better study actual decision-makers, rather than examining inputs and outputs. Finally, these findings force us to question whether conscription would serve as a check on foreign intervention (contra Haberman, 2017; Ricks, 2012), if casualties cause elites to escalate, rather than to end conflicts, a return to conscription may have the opposite effect from what its proponents champion.

Chapter 3: Identification, Interpretability, and Bayesian Word Embeddings

3.1 Introduction

Important questions in the social sciences turn on the meanings of words used to express ideas like language change, emotion, and ideological affinity (Hamilton, Leskovec, and Jurafsky, 2016; Pomeroy, Dasandi, and Mikhaylov, 2018; Rheault et al., 2016). One increasingly popular way to represent meaning, originating in natural language processing, is through the use of word embeddings. This class of models learns a set of coefficients which encode meaning by predicting a word given the surrounding words (Mikolov, Yih, and Zweig, 2013; Mikolov et al., 2013). These coefficients are the *embeddings*, which can then be used to analyze word meanings.

Unfortunately, existing embedding models are not always appropriate for answering social scientists' questions. Embeddings are not identified, and the dimensions are not directly interpretable, which makes it difficult to perform statistical inference on the embeddings produced by standard models, for example, using them as covariates in a regression model.[2]

[2]This is because in a regression setup, the coefficient is the change in the dependent variable for a 1-unit increase in the independent variable. With embedding dimensions, it is not clear what a 1-unit increase in the independent variable means, nor does direction have any clear meaning.

To resolve these issues, I cast word embeddings as a Bayesian latent variable model. Identifying multidimensional latent variable models is a known problem, and I draw on solutions proposed in the ideal point modeling literature (Clinton, Jackman, and Rivers, 2004; Rivers, 2003) to render embeddings interpretable and usable in a regression framework. I demonstrate these results on two corpora: a collection of inaugural addresses, and a selection of declassified diplomatic documents from the *Foreign Relation of the United States*. In the inaugural addresses, I find rhetoric became more domestically-focused after 1945, a shift which existing social science approaches cannot detect. This finding stands in contrast to what existing theories of international relations would have us expect. In the *FRUS* documents, I find that more bellicose rhetoric results in more aggressive American foreign policy behavior, helping confirm that elite deliberation matters for shaping foreign policy, and that the measurements I create correlate with existing datasets, helping to establish the validity of the model results.

3.2 Social Science and Embedding Models of Language

Traditional approaches to creating variables from text in the social sciences involve human coders, who assign documents to categories based on pre-defined criteria. However, this approach is expensive, and does not scale. Text as data techniques attempt to solve this problem through the use of natural language processing techniques to convert a corpus of text into numeric objects which makes inference possible (Gentzkow, Kelly, and Taddy, 2017; Grimmer and Stewart, 2013). These techniques allow scholars to create variables and operationalize concepts in corpora that are too

large for human coding, and investigate ideas which cannot be measured directly (unlike indicators like Gross Domestic Product or population).

While a variety of models have been proposed to create variables from political text, including scaling models (Lowe et al., 2011), and topic models (Blei, Ng, and Jordan, 2003; Grimmer, 2010; Roberts, Stewart, and Airoldi, 2016), these approaches focus on the document as the unit of analysis. Word embeddings, which have a long history in the natural language processing literature (see Turney and Pantel (2010) for an extensive review of pre-neural network models), have recently been embraced by social scientists for their potential for inference at the word level. Modern neural word embedding models learn a low-dimensional representation of a word as a dense vector by either factorizing a word co-occurrence matrix or predicting the co-occurrence of a pair of words using a single-layer neural network. Among the best known of these models is word2vec (Mikolov, Yih, and Zweig, 2013; Mikolov et al., 2013), which proposed an efficient model for learning embeddings, framing embedding learning as a prediction task, rather than a factorization task.

For social scientists, word embeddings are a powerful tool because they can represent the meanings of individual words. Embeddings can help isolate patterns in corpora that are expensive to label, and make apparent latent phenomena not observable through simple document-feature counts such as patterns of semantic change, (Hamilton, Leskovec, and Jurafsky, 2016), cultural assumptions and biases (Caliskan, Bryson, and Narayanan, 2017; Garg et al., 2018; Kozlowski, Taddy, and Evans, 2018), and ideological affinity in international organizations (Pomeroy, Dasandi, and Mikhaylov, 2018).

However, these embeddings can be problematic for social science research, where scholars care about both model identification and interpretable results. Embeddings are multidimensional latent variable models, which are not, by default, identified: a known problem with this class of model, where multiple permutations of latent dimensions can result in the same observed data (Aldrich, Montgomery, and Sparks, 2014; Clinton, Jackman, and Rivers, 2004; Rivers, 2003). However, by anchoring points on these dimensions, it is possible to present identified and interpretable dimensions. In the ideal-point literature, these anchors represent ideological "endpoints," with theory guiding the selection of which legislators are most liberal and conservative. Choosing words as anchors with a large number of dimensions is more difficult than choosing legislators. However, I offer a solution below.

There have been multiple efforts at developing Bayesian word embeddings (Barkan, 2017; Havrylov and Titov, 2018; Ji et al., 2017; Rudolph et al., 2016). However, none of these have exploited the key advantage of Bayesian inference: the ability to quantify the uncertainty in parameter estimates, and use prior information to inform parameter estimates. The one approach that has incorporated both uncertainty and hypothesis testing is Han et al. (2018), who offer both measures of uncertainty, and a way to test the effect of metadata on the similarity of embeddings. However, this approach does not account for identification problems in the learned embeddings.

3.3 Bayesian Word Embeddings

In this section, I develop word embeddings as Bayesian latent variable models estimated with variational inference, following similar work for probabilistic principal components analysis (Bishop, 1999) and ideal-point models (Imai, Lo, and Olmsted,

2016). I first discuss the embedding model setup, add Automatic Relevance Determination priors to the model, and then, present the variational updates to estimate the model.

Word embeddings predict the probability of a word-context pair co-occurring, and because the co-occurrence is a binary variable ($Y_{ij} = 1$ if w_i and w_j co-occur, 0 otherwise), I use a probit link function to model the probability of co-occurrence.

$$\begin{aligned} &p(Y_{ij} = 1) = \\ &(\mathbf{1}[z_{ij} > 0]\mathbf{1}[y_{ij} = 1] + \mathbf{1}[z_{ij} < 0]\mathbf{1}[y_{ij} = 0]) \\ &\mathcal{TN}(z_{ij}|\boldsymbol{x}_i^\top \boldsymbol{\beta}_j, 1). \end{aligned} \tag{3.1}$$

$\boldsymbol{X}$ and $\boldsymbol{\beta}$ are $K \times I$ (or $K \times J$, respectively) -dimensional matrices, $\boldsymbol{Y}$ is an $I \times J$ co-occurrence matrix, the corpus contains I words and J context words. Each embedding vector ($\boldsymbol{x}_i$ or $\boldsymbol{\beta}_j$) has a K-dimensional multivariate normal prior.

Most existing approaches to word embeddings contain no measures of uncertainty, or the covariance between dimensions. This can be a problem during estimation, as the model attempts to put equal weight on all dimensions. To resolve this, I use Automatic Relevance Determination (ARD) priors, which place a separate gamma-distributed scalar (e.g. α_{X_k}) on the diagonal for each dimension of the covariance matrix (Bishop, 1999; MacKay and Neal, 1994). These priors penalize unnecessary model dimensions, improving model fit.

This specification results in the following likelihood:

$$
\begin{aligned}
&p(\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\alpha}_{\mathbf{X}},\boldsymbol{\alpha}_{\beta}|\mathbf{Y}) \propto \\
&\quad (\mathbf{1}[z_{ij}>0]\mathbf{1}[y_{ij}=1]+\mathbf{1}[z_{ij}<0]\mathbf{1}[y_{ij}=0]) \\
&\quad \mathcal{TN}(z_{ij}|\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_j,1)\times \\
&\quad \prod_i \mathcal{MVN}(\boldsymbol{x}_i|0,\boldsymbol{\alpha}_X^{-1})\times \\
&\quad \prod_i \mathcal{MVN}(\boldsymbol{\beta}_j|0,\boldsymbol{\alpha}_{\beta}^{-1})\times \\
&\quad \prod_k \mathrm{Gam}(\alpha_{X_k}|c_{X_0},d_{X_0})\times \\
&\quad \prod_k \mathrm{Gam}(\alpha_{\beta_k}|c_{\beta_0},d_{\beta_0}).
\end{aligned}
\tag{3.2}
$$

For Bayesian models like this, the goal is to estimate posterior distributions of the parameters most likely to have produced the observed data. Given the joint density (probability of data and parameters), we want to calculate the conditional density of the parameters by evaluating the following integral (notation follows Bishop, 1999):

$$
P(\boldsymbol{Y}) = \int p(\boldsymbol{Y},\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3.3}
$$

where $\boldsymbol{\theta} = \{\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\alpha}_X,\boldsymbol{\alpha}_{\beta}\}$. This integral is analytically intractable, so we transform the integral using Jensen's inequality:

$$
\begin{aligned}
\ln P(\boldsymbol{Y}) &= \ln \int p(\boldsymbol{Y},\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \ln \int \mathcal{Q}(\boldsymbol{\theta})\frac{P(\boldsymbol{Y},\boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})}d\boldsymbol{\theta} \\
&\geq \int \mathcal{Q}(\boldsymbol{\theta})\ln\frac{P(\boldsymbol{Y},\boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})}d\boldsymbol{\theta} \\
&= \mathcal{L}(Q)
\end{aligned}
\tag{3.4}
$$

where $\mathcal{L}(Q)$ is evidence lower bound (ELBO).

The difference between the true model $P(Y)$ and variational approximation $\mathcal{L}(Q)$ can be represented as the Kullback-Leibler divergence:

$$KL(Q||P) = -\int \mathcal{Q}(\boldsymbol{\theta}) \ln \frac{P(Y|\boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{3.5}$$

so we turn to a mean-field variational approximation to estimate the model, minimizing the Kullback-Leibler divergence (Blei, Kucukelbir, and McAuliffe, 2017; Wainwright and Jordan, 2008). This requires assuming that the approximation to the posterior can be factorized:

$$\begin{gathered} \mathcal{Q}(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_\beta) = \\ \mathcal{Q}(\boldsymbol{Z}), \mathcal{Q}(\boldsymbol{X}), \mathcal{Q}(\boldsymbol{\beta}), \mathcal{Q}(\boldsymbol{\alpha}_X), \mathcal{Q}(\boldsymbol{\alpha}_\beta) \end{gathered} \tag{3.6}$$

and that appropriate approximating distributions can be found. In this case, the requirement is met: z_{ij} is approximated with a truncated normal, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_j$ are approximated with multivariate normals, and α_{X_k} and α_{β_k} are approximated with gamma distributions. This factorization and approximation can be further factorized

into the following parameter updates:

$$
\begin{aligned}
z_{ij}^* &= \mathbb{E}[\boldsymbol{x}_i^\top]\mathbb{E}[\boldsymbol{\beta}_j] \\
\mathbb{E}[q(z_{ij})] &= \begin{cases} z_{ij}^* + \frac{\phi(z_{ij}^*)}{\Phi(z_{ij}^*)} & \text{if } y_{ij} = 1 \\ z_{ij}^* - \frac{\phi(z_{ij}^*)}{1-\Phi(z_{ij}^*)} & \text{if } y_{ij} = 0 \end{cases} \\
\boldsymbol{A} &= \left(\text{diag}(\mathbb{E}[\boldsymbol{\alpha}_X])^{-1} + \sum_j \mathbb{E}[\boldsymbol{\beta}_j \boldsymbol{\beta}_j^\top] \right) \\
\boldsymbol{a}_i &= \sum_j \mathbb{E}[\boldsymbol{\beta}_j]\mathbb{E}[z_{ij}] \\
\mathbb{E}[q(\boldsymbol{x}_i)] &= \boldsymbol{A}^{-1}\boldsymbol{a}_i \\
\boldsymbol{B} &= \left(\text{diag}(\mathbb{E}[\boldsymbol{\alpha}_\beta])^{-1} + \sum_i \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] \right) \\
\boldsymbol{b}_j &= \sum_i \mathbb{E}[\boldsymbol{x}_i]\mathbb{E}[z_{ij}] \\
\mathbb{E}[q(\boldsymbol{\beta}_j])&= \boldsymbol{B}^{-1}\boldsymbol{b}_j \\
c_x &= c_{x_0} + \frac{I}{2} \\
d_{x_k} &= d_{x_0} + \frac{||\mathbb{E}[\boldsymbol{x}_k]||^2}{2} \\
\mathbb{E}[q(\alpha_{X_k})] &= \frac{c_x}{d_{x_k}} \\
c_\beta &= c_{\beta_0} + \frac{J}{2} \\
d_{\beta_k} &= d_{\beta_0} + \frac{||\mathbb{E}[\boldsymbol{\beta}_k]||^2}{2} \\
\mathbb{E}[q(\alpha_{\beta_k})] &= \frac{c_\beta}{d_{\beta_k}}
\end{aligned}
\tag{3.7}
$$

where $c_{x_0}, d_{x_0}, c_{\beta_0}, d_{\beta_0}$ are hyperparameters set by the user. Convergence is monitored via change in the ELBO, and when change drops below a user-specified threshold, the model is considered converged. This model is implemented in the R package `bwe`.[3]

3.4 Identifying Model Output

The output from multidimensional latent variable models is not identified, as many possible permutations of latent values can produce the same observed data

[3] https://github.com/adamlauretig/bwe.

(Rivers, 2003). However, by fixing $K(K+1)$ linearly independent values (anchors), users can guarantee the embedding matrix is identified (Bafumi et al., 2005; Clinton, Jackman, and Rivers, 2004; Rivers, 2003). To determine these anchors in the ideal point modeling literature, theory drives the endpoint selection: Clinton, Jackman, and Rivers (2004) fix both points for Jesse Helms, Ted Kennedy, and Lincoln Chaffee as right, left, and center anchors, respectively, in a $K = 2$ model.

While theory should always motivate modeling choices, determining theoretically motivated anchors when K ranges from 50 to 300 can be difficult. I propose a solution: theory can motivate initial anchor selection, and then, for each additional anchor, the most cosine dis-similar word is chosen as the opposite anchor. This allows the analyst to specify theoretically motivated opposites as initial anchors, and then, resulting anchors are chosen from remaining words. I provide an implementation of this algorithm in the R package `bwe`.

3.5 Interpreting Model Output

Anchoring the embeddings ensures they are identified. However, they are still not in a format which allows for ready interpretability in the regression-based models social scientists are most familiar with. To transform embeddings so that they can be used in regression, I opt for a modification of the anchoring approach discussed above. For this approach, the user specifies a pair of endpoints for a dimension, where the endpoints of interest are set to 1 and -1. This can be applied to as many dimensions as necessary, and then the automatic, cosine-based anchoring is used for the rest of the dimensions. An affine transformation is then used to transform the embedding matrix relative to the chosen anchors.

A key advantage of this approach is that because two anchors are supplied, words are scaled on this dimension. For example, while simply choosing "war" as an anchor results in the results words scaled according to their similarity with "war", setting "war" and "peace" as opposite anchors (1 and -1, respectively) allows for a measure of bellicosity in a corpus.

This method can be applied to as many words/concepts as the user is interested in (as the automated cosine similarity will handle the other dimensions), and, of note for social scientists, each of these word scalings, which are $I \times 1$, can be multiplied by a $D \times I$ document-term matrix: $D \times I * I \times 1$, scaling the documents in a corpus according to dimensions of interest. These document values can then we used in a regression, and the coefficients can be interpreted in a straightforward way.

3.6 Inaugurals and Internationalism

In an initial test of this model, I investigate whether the United States saw itself in a new, global role after 1945, as perceived in presidents' inaugural addresses. After 1945, the United States was the global hegemon, and international relations theory argues that this resulted in a shift in American attitudes towards the world (Mearsheimer, 2001). It has been shown that the public takes elite cues on various issues (Druckman and Jacobs, 2015; Zaller, 1992), and since foreign-policy is generally viewed as an elite-led phenomenon (Aldrich et al., 2006), I explore whether, after the second World War, inaugural addresses were more internationally focused than those before the war.

I use the corpus of inaugural addresses available in the `quanteda` R package (Benoit et al., 2018), which contains 58 speeches. I keep words which occur with

frequency > 5, and then lowercase and tokenize the texts, resulting in 2705 words. I estimate the model with a context window of 9, with 5 negative samples for every positive sample, and the number of dimensions $K = 50$. After fitting the model, I compare three possible anchorings: an un-anchored embedding, an embedding anchored on "american," and an embedding scaled with the first dimension anchored on "international" and "domestic;" the results are visible in Table 3.1. We see that changing the anchoring points changes the most similar words. However, anchoring helps make these embeddings more interpretable. To test whether there was a statistically significant difference between American perceptions of global roles before and after 1945, I multiply the document-term matrix by the embedding dimension anchored on "international" and "domestic," creating an "internationalism" scale for documents. I test this hypothesis using a one-sided Kolmogorov-Smirnov test, and reject the null hypothesis, that pre-1945 inaugural addresses are less internationalist than the post-1945 addresses at $p < .05$. This means pre-1945 addresses are more "internationalist" than the post-1945 addresses. I plot the differing distributions in Figure 3.1.

What explains this finding? One possibility, building on Herring (2008, ch. 1), is that the United States was not isolationist prior to 1945, that isolationism was largely a product of the 1920s and 1930s. However, the United States was more unilateral before 1945. Because the multilateral world order was a fact of life after 1945, it is possible presidents were less likely to comment on international affairs, international action was the norm, rather than the exception. Furthermore, the public played a larger role in shaping foreign policy action, particularly during the Vietnam War era,

Word of Interest	Anchor: International, Domestic							
war	large	declarations	pay	carefully	choice	equal	this	greater
peace	practices	meeting	strife	inspiring	confederacy	advance	temple	objections
american	engagements	soil	cultivate	by	heroes	goes	pride	she
international	declare	path	honor	expression	speaking	where	vision	ignorance
national	temple	subject	learned	demand	advance	objections	principle	guard
	Anchor: American							
war	abroad	remedies	violate	slaves	violence	declarations	proposition	sectional
peace	army	plenty	victory	effort	resumption	front	regulation	agreement
american	regards	brief	instrumentality	execute	able	friendly	hands	friendship
international	assembly	european	continent	capable	various	canal	differing	affected
national	now	recognition	corporations	monetary	south	more	character	diversity
	Anchor: None							
war	made	had	peace	force	never	after	still	place
peace	world	nations	war	strength	prosperity	progress	just	security
american	through	opportunity	america	life	justice	right	individual	equal
international	maintain	lasting	fixed	beneficial	settlement	likely	relationship	intercourse
national	most	necessity	common	given	free	first	an	power

Table 3.1: The most similar words to "war," "peace," "american," "international," and "national," according to each of the anchoring choices, measured via cosine similarity. Choosing appropriate anchors leads to more interpretable embeddings than the unanchored model.

than it had previously (Aldrich et al., 2006), and this could lead to a blurring of the lines between foreign and domestic politics when presidents address the public.

I compare the results from Bayesian Word Embeddings to the results from a standard model used in the social sciences to analyze text, the structural topic model (Roberts, Stewart, and Airoldi, 2016)[4]. I find that "domestic" and "international" topics are not linked, the structural topic model captures no relationship between these words. I then investigate the change in "domestic" and "international" topics before and after 1945, and find no effect. There is ample belief and qualitative evidence of a change in American views about the world after 1945, which is not captured in the structural topic model. These differing results suggest that the embeddings are capable of recovering patterns in language that document-based topic models cannot.

[4]Results presented in the Appendix.

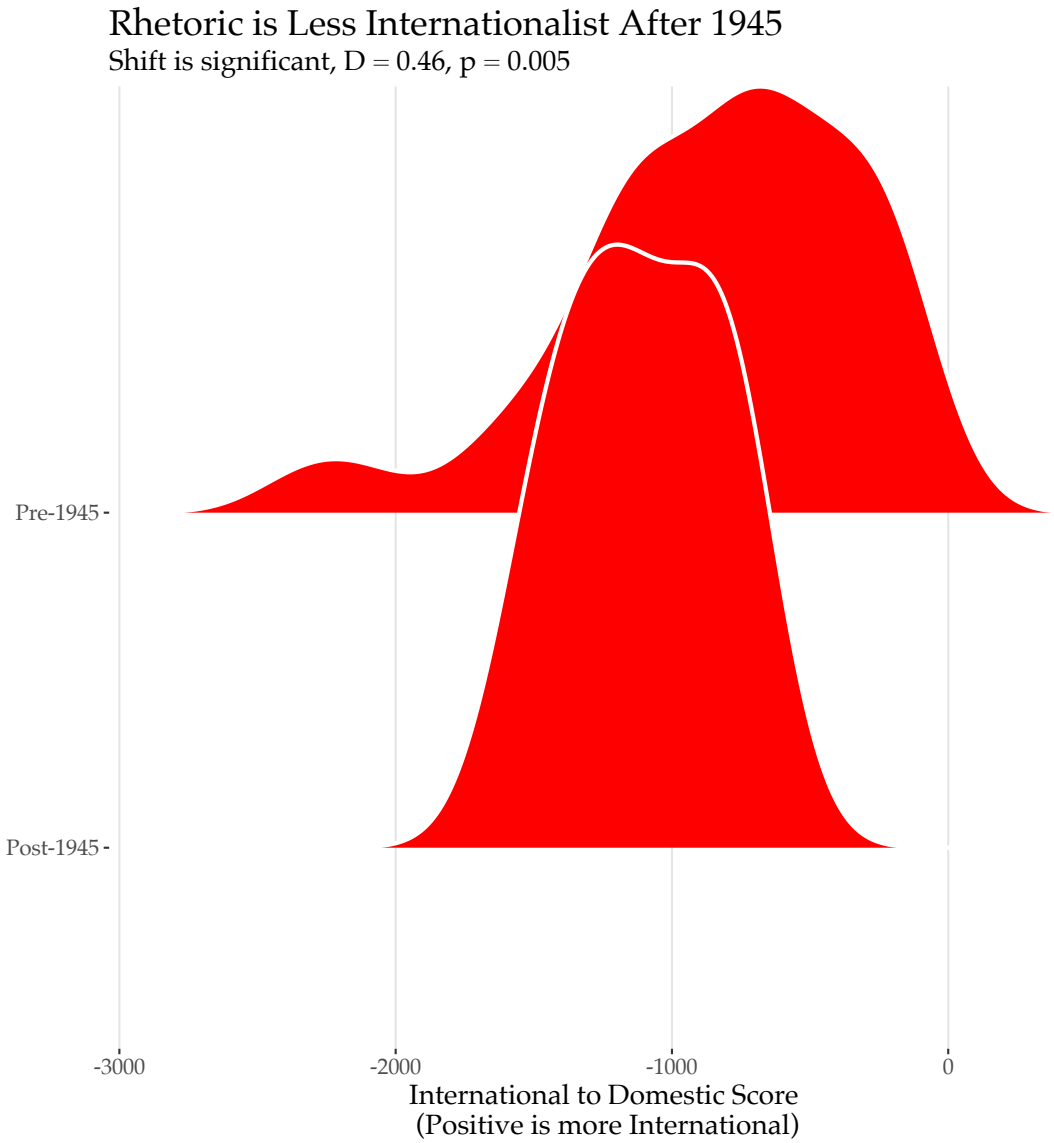

Figure 3.1: After 1945, rhetoric in inaugural addresses becomes less internationalist, and more domestic.

3.7 Diplomacy and the Onset of War

Natural language processing and text as data methods offer the opportunity to quantify decision-making and attitude among elites, which is notoriously difficult to measure, especially in times of conflict. Existing approaches to measuring elite attitudes often depend on survey or laboratory experiments (Feaver and Gelpi, 2006; LeVeck et al., 2014). However, I offer an alternative approach that allows us to examine elite decisions as they occur. Drawing on a novel corpus of recently digitized diplomatic cables, the *Foreign Relations of the United States* (FRUS), I investigate whether changes in the bellicosity of elite rhetoric precedes an escalation in US hostility. The FRUS dataset provides an exciting opportunity to investigate bellicosity among American foreign policy elites as events happened, as it contains primary source documents of private communications from the policymakers who develop and implement the United States' foreign policy. Among the sources for documents included in FRUS are "Presidential libraries, Departments of State and Defense, National Security Council, Central Intelligence Agency, Agency for International Development, and other foreign affairs agencies as well as the private papers of individuals involved in formulating U.S. foreign policy," with a focus on documents relevant to policy-making (State Department, 2017). When a FRUS volume is compiled, the compiler(s) first identify a set of themes, develop a list all relevant documents, and then select those with the greatest historical import. These are then redacted or declassified, typeset, compared to the original document, and printed and bound (McAllister et al., 2015).

To explore elite bellicosity, I investigate behavior during 1964-1966, the leadup to the Vietnam War, and the breakdown of the "Cold War Consensus" (Krebs, 2015).

The era is particularly interesting because, while the United States increased its commitment to Vietnam, it also engaged in several other interventions around the world (Herring, 2008, ch. 16). Thus, we would expect to see that an increase in bellicosity in the FRUS corpus would be correlated with an increase in hostile actions by the United States.

I measure hostility using the Cline Center Historical Event Data, coded from the *New York Times* (Althaus et al., 2017). These data take the form *(DATE, STATE A, ACTION, STATE B)*, where *(STATE A, STATE B)* are directed dyads, *DATE* is the day the event was observed, and *ACTION* is one of five categories of action: neutral, verbal cooperation, verbal conflict, material cooperation, or material conflict (Norris, Schrodt, and Beieler, 2017). I select only those events where *STATE A* is the United States, and sum events at the biweekly level. I measure hostility using counts of material conflict events, and display the hostile event counts in Figure 3.2.

To calculate bellicosity, I first estimate a Bayesian Word Embedding model, with context window of 9, $K = 50$, keeping any word that occurs at least 40 times. I then anchor the embeddings on a “war-peace" dimension. I summarize the results of the anchoring using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), which calculates a low dimensional number of components, similar to principal components analysis. Unlike PCA, UMAP calculates distance using cosine similarity, while balancing both global and local structure in the embeddings (Becht et al., 2019). I present results in Table 3.2, and the components reveal themes in the corpus, clustered by region and issue, helping highlight the face validity of the embeddings.

iran	doubtful	communications	tam	bases	relatively	robertson	initials
iranian	blocked	relations	systematically	family	zambia	outflows	secretary
shah	sponsored	appreciably	north	leave	udi	payments	footnotes
aram	ultimatum	masses	hanoi	deployments	tran	liabilities	present
iranians	telecommunications	sites	recce	fixed	rhodesia	fowler	conflict
afghan	recommendation	overtures	drv	precondition	sr	banks	president
squadron	imminent	concurs	vinh	laotian	neighboring	tax	raymond
mnd	jet	harass	chau	reasons	continent	corporations	even

Table 3.2: The top words from a subset of components estimated from UMAP. Components include a variety of regional and substantive themes. These results help highlight the validity of the embeddings: semantically similar words are appearing near each other in cosine space.

To estimate the bellicosity of a given document, I multiply the war-peace dimension by the document term matrix, averaging document bellicosity scores at the bi-weekly level. I plot the bi-weekly bellicosity scores in Figure 3.2.

To determine if there is a relationship between hostile events and bellicosity, I regress events on the lagged bellicosity (to account for a delay in policy implementation), using a Poisson generalized linear model, due to the count-distributed nature of the outcome.[5] I plot the regression line against the data in Figure 3.3, and find a positive and statistically significant effect.

This result suggests that bellicosity in elite deliberations, captured in diplomatic documents, results in an increase in conflictual events, which suggests that the documents in the *FRUS* corpus do not simply contain cheap talk, these deliberations ultimately shape policy. These findings also help establish the validity of the "bellicosity" scale, that is, it correlates with an entirely separate dataset, which captures a similar phenomena. All replication materials are available at `https://github.com/adamlauretig/bwe_application_naacl_2019`.

[5]In the appendix, I remove outliers and high-leverage points from the dataset, and fit the model again. Results do not change.

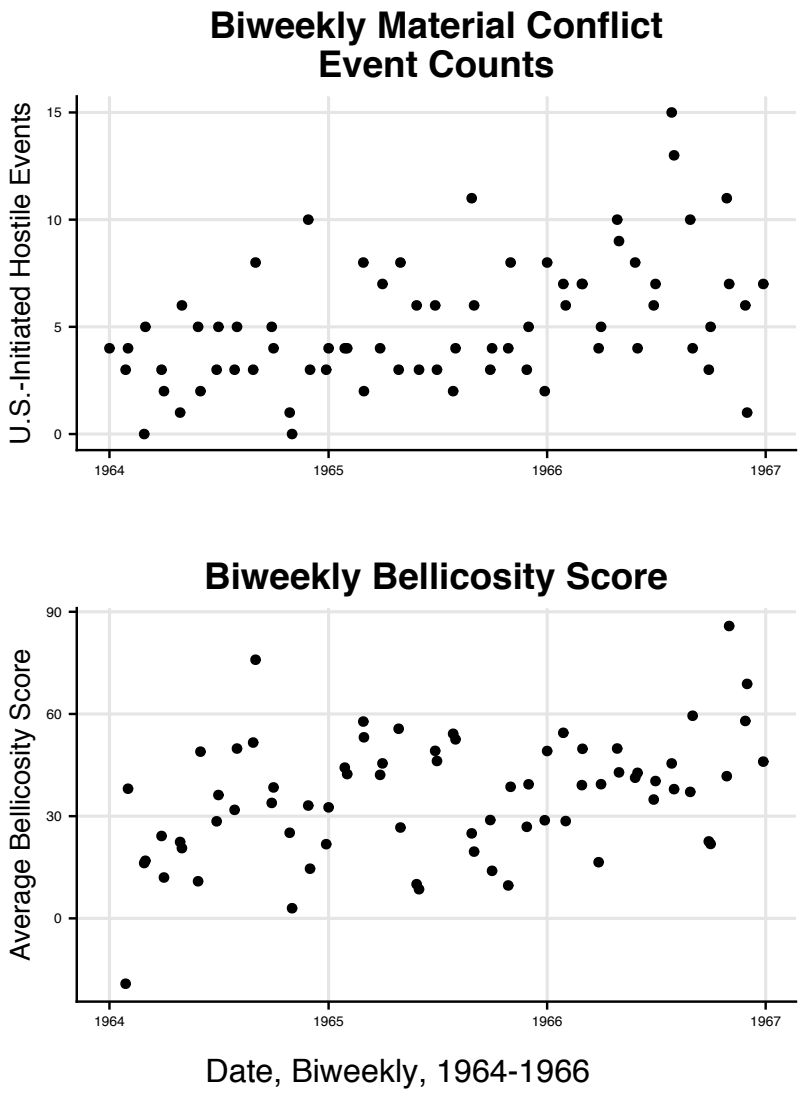

Figure 3.2: Plotting material conflict event counts and bellicosity scores over time aggregated at the bi-weekly level. Both bellicosity and the count of material conflict events increase with time, as the United States increased its involvement in the Vietnam War.

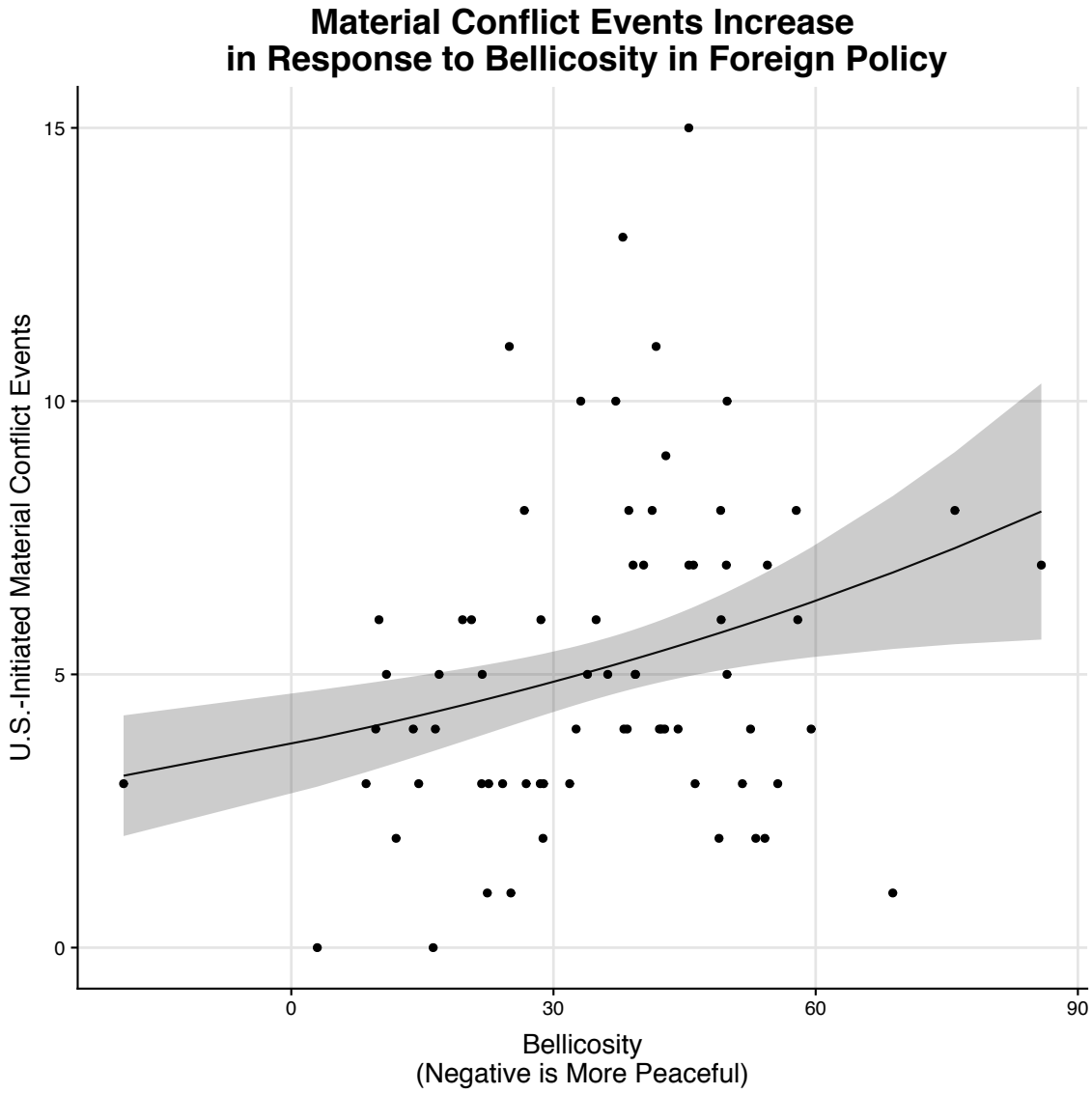

Figure 3.3: An increase in the previous bi-weekly period's bellicosity is associated with an increase in U.S.-initiated hostile events. The regression is from a Poisson generalized linear model, and uncertainty is displayed with 95% confidence intervals.

3.8 Conclusion

In this paper, I introduced Bayesian Word Embeddings, a method for estimating word embeddings which uses variational bayesian methods. I incorporated Automatic

Relevance Determination priors on the embedding dimensions, relaxing the requirement that all dimensions have equal weight. Linking word embeddings to Bayesian latent variable models, I then discussed issues with identification, and solutions proposed in the ideal-point literature, as well as offering an alternative which allows for scaling along dimensions of interest, which creates model that can then be used in a regression.

I applied Bayesian Word Embeddings to two cases: examining the change in American attitudes about the world before and after 1945 as captured in Presidential inaugural addresses, and then, testing whether an increase in the bellicosity of internal elite discussion (in diplomatic documents) results in an increase in American hostility. I found that there was a statistically significant different in the views of the world expressed in inaugural addresses, and that this shift was the opposite of what hypotheses generated from international relations theory would expect. When testing the effect of bellicosity on the hostility of American foreign policy, I show that an increase in bellicosity resulted in an increase in hostility.

Overall, I have contributed a tool which can serve many purposes for social scientists. By building a probabilistic embedding model, I have constructed a tool which moves beyond document-based inferential approaches to text as data, allowing inference on individual words. This promises new reaches for social scientists, in particular, the promise of crossover with interpretivist work, building on Nelson (2017). Concepts such as securitization theory (Wæver, 1995) draw on the idea that language and word choice by elites shape the attitude of the public, and through the methods introduced above provide the opportunity to generate statistical tests for hypotheses derived from theories like securitization theory.

Future methodological work will follow three tracks. The first will build on Rudolph et al. (2016) and Han et al. (2018), one goal is incorporating document-level metadata into embedding estimation, allowing embeddings to vary according to document-specific attributes, and then, identifying the resulting embeddings. The second will take advantage of stochastic variational inference (Hoffman et al., 2013) to enable Bayesian Word Embeddings to scale to massive corpora. Finally, the third track for future work will involve tying the anchoring approach discussed above with the emerging literature on making casual claims from text (Fong and Grimmer, 2016; Mozer et al., 2018), and taking advantage of the word similarities to identify appropriate linguistic counterfactuals.

Chapter 4: Bayesian Word Embeddings, Securitization, and U.S. Foreign Policy

4.1 Introduction

What motivates foreign policy decision-making? Scholars have debated this extensively, with some arguing that behavior is rooted in material capabilities, and others, that ideas motivate foreign policy decision-making (Brooks and Wohlforth, 2001). One problem with many of these debates is that the mechanisms by which decisions are made are often overlooked. In particular, there has been a dearth of quantitative tests of cases where ideational factors shape decision-making.

One mechanism that remains under-explored in the quantitative international relations literature, despite its prominence in the qualitative literature, is securitization theory. Securitization theory is built around the idea of the *speech act*, that discourse shapes attitudes around foreign policy issues, and thus, the actions states will take. In particular, in securitization, an issue is moved from the realm of "ordinary politics" to "extraordinary politics," a state of emergency where normal rules no longer apply, and issues are elevated to the same severity as physical security threats (Buzan, Wæver, and De Wilde, 1998; Wæver, 1995).

In many cases securitization is initiated by elites through declarations which aim to change the set of actions available around an issue. In this way, securitization can

be thought of as a case of elite-cue theories of political behavior, where the cues are verbal or textual (Zaller, 1992). Quantitative modeling of elite behaviors is difficult, especially if one cannot get access to elites for experiments (unlike Feaver and Gelpi, 2006 or LeVeck et al., 2014). Accessing the text of policymakers' deliberations as the securitization process unfolds is particularly difficult, which limits scholars' ability to determine what affects elite attitudes, and how elite attitudes can result in policy changes.

To investigate the determinants and effects of securitization, I use the recently digitized *Foreign Relations of the United States* corpus, a collection of primary-source diplomatic documents, including such sensitive documents as Henry Kissinger's back-channel communications. I use the Bayesian Word Embedding model and scaling technique developed by Lauretig (2019) to model the meanings of words, and scale documents according to their bellicosity, which I take as a measure of securitization. I then regress measures of foreign policy behavior on this bellicosity measure.

I test securitization across three cases: one where ideology could motivate securitization, one where material capabilities could motivate securitization, and one where both play a role. These cases are the Cuban Revolution in 1959, France's development of a nuclear bomb in 1960, and China's development of a nuclear bomb in 1964. Surprisingly, I find that securitization has no effect on the American policies directed at each of these three countries, regardless of whether these are cooperative or conflictual policies. This is despite the fact that ideology increases the securitization of relations between countries, but material capability does not, in keeping with what prior knowledge would have us expect. I conclude with some thoughts for why this

null finding may be the case, and the implications for the study of foreign policy decision-making.

4.2 Overview of Securitization Theory

Securitization theory is a theory about the way in which elites discuss policy issues. The core of securitization theory is that elites move certain issues into the realm of exceptional politics, declaring these issues an existential threat, on par with physical security issues. Securitization is based on the analysis of language and process, such as the declarations of policymakers about what is being threatened, by whom, and the audience the policymaker seeks to persuade (Buzan, Wæver, and De Wilde, 1998; Wæver, 1995).

Traditional realist approaches to security treated security as "objective" and as a given, threat was simply the product of a combination of material factors. For example, during the Cold War nuclear weapons were considered a security threat, regardless of who held them (Jervis, 1978). As constructivism developed in the 1990s, one of the key insights was that security was not objective, that it was mediated by perception and the social construction of threat. As Wendt (1999, p. 255) noted, not all nuclear weapons were equal: "five hundred British nuclear weapons are less threatening to the US than five North Korean ones because of the shared understandings that underpin them." Threat, danger, or security are not simply a function of material capabilities, but rather, material capabilities are understood through a lens of ideas, interests, and perceptions. While constructivism highlighted the importance of ideas and perceptions in understanding behavior in international politics, it did

not provide for much in the way of mechanisms to explain *why* ideas and perceptions of threat can change.

Securitization theory, or the Copenhagen school (Wæver, 1995) provides a plausible mechanism to mediate the relationship between material capabilities and perceptions of threat. Securitization theory developed after the Cold War as a reaction to the realist/materialist-driven approaches to security which dominated during the Cold War, and it emphasizes the *process* by which entities are moved into the realm of existential threat, rejecting the idea that any entity is inherently threatening. Balzacq (2010, p. 3) highlights three factors of securitization theory to understand the process of securitization: "the centrality of the audience, the codependency of agency and context, and [...] a constellation of practices and tools." Central to securitization is the interaction between the policymaker making securitizing claims and the audience, who receives them. This is because securitization is rooted in *speech acts*, which are comprised of locutionary, illocutionary, and perlocutionary acts (Austin, 1975). These acts are about what is said, what is meant, and what is done in communication. The perlocutionary act of securitization requires the audience to accept the securitization frame: while elites craft a give speeches and discuss an issue (locution), in order to move an issue into the realm of security (illocution), the securitization act is only complete when the audience accepts the movement into the realm of security. One could pithily summarize this by asking "what if they tried to securitize an issue and nobody cared?" The answer: it would not be securitized.

In many ways, securitization is the critical theory equivalent to the elite-cue theory of public opinion (Berinsky, 2007; Druckman and Jacobs, 2015; Zaller, 1992). Much like the elite-cue literature, securitization proposes that elites are the source of much

of public opinion. What constitutes an elite cue varies across the literature, with Berinsky (2007) casting it as dissent over war, and Druckman and Jacobs (2015) as a concerted effort by presidents to manipulate public opinion. To understand securitization in this framework simply requires casting elite cues as speech acts, that is, the impetus for changes in public opinion are the communications from elites. The difference between elite-cue theory and securitization theory is the speech-act component: the actions taken as a result of securitization. For example, Mueller and Stewart (2012) could be presented as an example of securitization, using the language of mainstream political science: policymakers cast terrorism as an existential threat, this threat is then used to justify a variety of extreme measures.

Securitization theory has traditionally been applied to a variety of issues which are not traditionally considered security issues, in order to understand how they became seen as emergency issues. Two examples of this are analyses of environmental issues and migration, which have played central roles in the post-Cold War security discourse. Much of the original work on securitization after the Cold War was developed to understand environmentalism, and the treatment of environmental issues as an existential risk. Linking concerns about the environment with national security was seen as a risk by those concerned about environmental issues, as the zero-sum logic of national security, as applied in interstate violence, could undermine efforts at international cooperation to solve climate issues (Deudney, 1990). While securitizing the environment can lead to the allocation of additional resources to tackle problems, it also places environmentalism in the realm of exceptional politics, and removes many options for tackling issues, options which can be preserved if environmentalism is simply cast as "daily politics" (Graeger, 1996).

Migration is the other major subject in the post-Cold War era to which securitization theory has been applied, to understand why migration is viewed as a "threat." Migration is framed as *destabilizing*, something which will upend the domestic order of a state, with transnational communities potentially undermining the nation-state (King and Melvin, 2000). In particular this framing has been applied in Europe, analyzing the contrast between the Schengen zone and other countries (Huysmans, 2000). Ibrahim (2005) notes that this securitization often has a racialized element: migrants from non-western countries are often framed as threats.

While securitization theory is often used to understand how non-military issues become securitized, it can also be applied to understand military issues, moving beyond the crude comparisons of material capabilities endemic to the realist literature. While material capabilities matter, Buzan (2008) highlights how many of the power dynamics realists often attribute to material capabilities are often explained by securitization, and the process by which certain states become threats, and others do not. For example, while the Cold War was characterized by a bipolar alliance structure, and the embrace of nuclear weapons, Buzan (2004) argues that the ideological divisions of the Cold War led to its securitization. Unlike traditional great power rivalries, the Cold War was cast in zero-sum ideological terms, with the United States and Soviet Union each casting their ideological vision as the only possibility for humanity as a whole. This ideological shift is the securitizing element of the Cold War: elites framed the issue as an existential crisis, that must be dealt with using extraordinary powers. Furthermore, nuclear weapons were themselves securitized, treated as exceptional and cataclysmic weapons which could destroy European civilization (Buzan

and Herring, 1998). Securitizing nuclear weapons also led to the rise of nonproliferation regimes, as concerns about the deadly capabilities of these weapons provoked a backlash among civil society groups.

4.3 Testing Securitization Theory

One major gap in our understanding of securitization theory at this point is an understanding of the effects of securitization on outcomes quantitative scholars care about. Securitization is treated as a process. However, there is a paucity of understanding what initiates the securitization process around a particular issue. There has been little existing effort to construct quantitative tests of securitization theory (excepting the framing experiment in Caverley and Krupnikov, 2017), with securitization existing parallel to much of the standard quantitative scholarship on foreign policy decision-making. This bifurcation develops from the conflation of "positivist" and "quantitative," usually by intepretivist scholars. The interpretivist disdain for quantitative work emerges from the thinking that interpretation and perspective can only be accomplished via qualitative work (Aradau and Huysmans, 2014; Salter and Mutlu, 2013). There has been pushback on this point, with Barkin and Sjoberg (2015) arguing that the epistemology of an argument is separate from the methodology used to answer a question: positivism and interpretivism epistemologies are orthogonal to quantitative and qualitative methodologies. In particular, there is nothing about quantitative methods which make them inherently positivist, nor qualitative methods which make them inherently interpretivist (Barkin and Sjoberg, 2015). For example, survey data about national identity can be used compressed with principal components analysis to create measures of identity (Kowert, 2017), which allow the approach

developed by Hopf (2002) to be scaled, and compared across countries. However, this combination of interpretivist thought and statistical techniques is still in the minority.

If interpretivist scholars are willing to take advantage of modern machine learning methods to scale their analyses (Gerring, 2017; Nelson, 2017), they can extend their inferences beyond single-case studies, and, solve two problems which commonly arise when trying to link securitization theory with mainstream political science: measurement and inference. The first problem for conventional political scientists who want to investigate securitization is the difficulty of operationalizing securitization. Despite the fact that securitization is presented as a continuum theoretically, in practice, securitization is often treated as a binary: either an issue is securitized, or it is not. This binarization is problematic because an issue may be securitized, but not every issue is securitized to the same degree. This suggests that there needs to be a scale for securitization, where one end of a scale is "securitized" and the other is "desecuritized."

The second problem with securitization theory is that it is never presented with a counterfactual, either as the independent or dependent variable. While some have discussed "desecuritization," where an issue is moved out of the realm of extraordinary politics, and back into the realm of "normal" politics, as the opposite of "securitization," there is little discussion of *why* an issue gets desecuritized (Wæver, 1995). Securitization is presented as a process. However, there is almost never neither an inciting incident, nor an outcome. This can be frustrating to researchers accustomed to a world of testable hypotheses. This is related to the problem of measurement, as without a clear defining line for when securitization "begins," it is impossible to falsify hypotheses about securitizing behavior.

To resolve these problems, I treat securitization as an independent variable to investigate the effects of securitization on foreign policy actions: if a relationship between the United States and another country becomes more materially hostile after the relationship between them becomes securitized in three cases. These three cases all have discrete events, where history and qualitative research leads us to believe there is a clear change in either ideological or material conditions which could prompt securitization.

4.4 Cases

I use three cases to investigate the effects of securitization: the Cuban Revolution, the French development of a nuclear bomb, and the Chinese development of a nuclear bomb. Here there is variational in both material capabilities (the development of nuclear weapons) and ideology (Communism). Based on the historical record, these would be ideal cases to investigate securitization, due to the sudden change which could have prompted securitization.

4.4.1 Cuba

The Cuban revolution represents a change in ideological orientation of what had previously been a major American ally, and plausible first case for the ideological drivers of securitization theory. It also presents the clearest anecdotal evidence for securitization, as Cuba has been cast as a danger to the United States since 1959 (DeYoung, 2014), despite being a poor and small island. Here, if ideology drives securitization, we would expect to see an increasingly hostile view of Cuba after the Revolution in 1959.

After the Spanish-American war, Cuba was an American protectorate, ruled by a series of American-backed strongmen, the last of whom was Fulgencio Batista. Batista was overthrown by a revolutionary named Fidel Castro in 1959. While initially the United States did not take Castro seriously, and even sent diplomats to Cuba to engage with him, this soon shifted drastically (Paterson, 1995). Castro legalized the Communist party, brought leftists into the government, and most outrageous to the United States, sought a trade deal with the Soviet Union (Herring, 2008, p. 688).

The Eisenhower administration decided that Castro had to go, a decision which set the tone for the next 60 years of American-Cuban relations. Given the proximity of Cuba to the United States, and Castro's Communist-influenced nationalism, which inspired other revolutionaries in South America, the United States viewed Castro as an existential threat (Rabe, 1988).

This ultimately led to an aggressive foreign policy of extended sanctions, as well as direct action to remove Fidel Castro, most famously, the Bay of Pigs catastrophe (Higgins, 1987). American-Cuban relations ebbed and flowed over the next 60 years, with the Cuban missile crisis representing a definite low point. However, Cuba was consistently treated as an existential threat by American political leaders, even after the end of the Cold War.

4.4.2 France

Unlike the Cuban case, French nuclearization represents a close ally fundamentally changing the balance of material capabilities, which represents a test of the realist case that material capabilities dictate the relationships between states. The French nuclear program represents an interesting test case for securitization: a close

ally of the United States acquired nuclear weapons, which in realist theory, would fundamentally change the relationship between the United States and France (Sagan, 1997). If material capabilities drive securitization, we would expect French acquisition of nuclear weapons to lead to an increasingly hostile view of France by American policy-makers.

France acquired nuclear weapons in 1960, a process which began in the mid-1950s. Realists cast this as a security-motivated decision, claiming that after the 1956 Suez crisis, France sought a nuclear weapon to ensure that it could continue to act as a great power in the international system without being coerced by others. However, see Sagan (1997) for a debunking of this logic.

French acquired nuclear capabilities not for security reasons, but rather reasons of prestige. After World War Two, while France was on the winning side, its capabilities and prestige were both severely diminished, and it lost control of its two largest colonies Indochina (Vietnam) and Algeria, which significantly undermined France's view of itself as a great power (Porch, 2013, ch. 5).

Charles de Gaulle explicitly centered France's nuclear program in a desire for France to be recognized as a great power (Kohl, 2015, ch. 4). To De Gaulle, sovereignty was synonymous with nuclear capability (Trachtenberg, 2012, p. 88), this desire for nuclear capability was not motivated by any particular threat. De Gaulle saw a future for a French-led Europe balancing against both the United States and Soviet Union, but the nuclear weapons which would make this possible were exclusively *French*, not European. In particular, De Gaulle did not want to share nuclear weapons with the Germans, which was understandable, given his memory of the Second World War (Trachtenberg, 2012). After nuclearization, de Gaulle flexed his

diplomatic muscles, withdrawing from NATO, and demanding that NATO troops be removed from France, and the United States acquiesced. While the United States was not happy about this, and there was serious consideration of cutting off funding to Europe, French nuclear capability was not cast as an existential risk to American interests (Herring, 2008, p. 743).

4.4.3 China

The third case of securitization I examine is Chinese nuclearization. China was already seen as hostile to the United States, and then, in 1964, China developed a nuclear bomb. Here, then, we can see whether a change in material capabilities affects the securitization of a relationship, when a country is already perceived as a threat, rather than an ally, as in the case of France. We might expect that China is seen as more hostile than France, after nuclearization, because China was a Communist power before nuclearization.

After the Chinese Revolution in 1949, the United States refused to recognize the Communist government, considering Taiwan to be the "true" Chinese government (Herring, 2008, p. 637). After the First Taiwan Straits crisis, China, under Mao Zedong, began developing nuclear weapons, to gain diplomatic credibility in the eyes of the United States. Mao sought nuclear weapons strictly as a deterrent, not as offensive weapons.[6] When the United States learned of the Chinese goals for a nuclear bomb, the Kennedy administration considered this an unacceptable risk. In particular, policymakers in the Kennedy were concerned that Chinese acquisition of a nuclear bomb would weaken American influence in Asia. Chinese nuclear capability

[6]This is the security dilemma (Jervis, 1978). Conveying intent is difficult, especially the intent that nuclear weapons would not be used offensively.

would pay "psychological dividends," casting Communism, and the Chinese model for growth and industrialization, as the wave of the future (Burr and Richelson, 2001). This ties into the notion of securitization and the Cold War that Buzan (2008) notes: the stakes were not simply over territory or resources, but rather, a zero-sum competition for the future of humanity. Then-President Kennedy considered a variety of options to slow or halt China's nuclear program, including cooperation with the Soviet Union. Other options considered included "infiltration, sabotage, or invasion by Chinese Nationalists; maritime blockades; a South Korean invasion of North Korea (to put pressure on China's borders); conventional air attacks on nuclear facilities; and the use of a tactical nuclear weapon" against China (Burr and Richelson, 2001, p. 68). That these measures were considered highlights how American policymakers considered a nuclear China an existential risk, and that the risk came from the interaction between nuclear capability and ideology.

After China tested a nuclear bomb in October 1964, concern about Chinese nuclear capabilities continued. Rather than engage in military action, however, the Johnson administration pursued a policy of "export controls, military containment, and continued intelligence monitoring," rather than the military actions Kennedy fancied (Burr and Richelson, 2001, p. 96). However, the sustained anti-nuclear China stance could not hold, and ultimately the United States acquiesced to a nuclear China, desecuritizing their relationship, and no longer treating China as an existential threat, the process of desecuritization that Wæver (1995) discussed.

4.5 Data

Measuring securitization requires access to a text of elite discussions, a set of securitizing events, and an outcome. For elite discussion, I use a subset of the *Foreign Relations of the United States* (FRUS) corpus, for a set of securitizing events, I hand-code indicator variables, and for an outcome, I use the Cline Center event data (Althaus et al., 2017). I detail the FRUS corpus and the Cline Center event data below.

4.5.1 FRUS Corpus

The FRUS corpus contains primary-source records of communications between elites, which allows scholars to observe elite deliberations as they occurred. It can be difficult to access elite opinions and attitudes, and even more difficult to quantify them; many existing efforts to measure elite attitudes resort to survey or lab experiments (Feaver and Gelpi, 2006; LeVeck et al., 2014). The FRUS corpus, combined with text-as-data and natural language processing techniques, offers a way to measure decision-making in real time. These documents capture the private communications of elites, and so, if we want to examine the processes by which decisions occur, and these documents do not illustrate this process, nothing would. The FRUS corpus contains documents from "Presidential libraries, Departments of State and Defense, National Security Council, Central Intelligence Agency, Agency for International Development, and other foreign affairs agencies as well as the private papers of individuals involved in formulating U.S. foreign policy," with a focus on documents relevant to policy-making (State Department, 2017). When a FRUS volume is compiled, the compiler(s) first identify a set of themes, develop a list of all relevant

documents, and then select those with the greatest historical import. These are then redacted or declassified, typeset, compared to the original document, and printed and bound (McAllister et al., 2015).

I use all FRUS documents from 1957-1976, 47964 documents[7]. I retain all words with frequency > 40, which are more than three characters long, and I remove common stopwords, roman numerals, and `XML` tags, leaving 13121 words. I display some corpus characteristics in Figure 4.1, which highlights that there is a frequency/length tradeoff in the corpus. If there are more documents in a given month, they are shorter, and if there are fewer, they are longer. I then plot the most frequent words in Figure 4.2, and the top words are unsurprising: a combination of political leadership words ("president", "secretary"), political organizations ("military", "soviet"), and functional words ("said", "time", "also"). I then use regular expressions to extract document types from document metadata, and present the twenty most common types in Figure 4.3. These are generally documents used to communicate between two parties ("Telegram/Memorandum/Letter from"), and the corpus also includes the backchannel messages used by Kissinger to circumvent standard State Department communications, suggesting that these documents do include communications which capture the actual policy deliberations of elites, not simply empty platitudes.

4.5.2 Cline Center event data

To measure whether securitization affects the behavior of states, we need an outcome, focused on *actions*. To measure foreign policy actions, I turn to the Cline Center Historical Event Data, coded from the *New York Times* (Althaus et al., 2017). These

[7]Prior to 1957, dates are inconsistently recorded in document metadata, and after 1976, documents are still being declassified, leading to inconsistent regional coverage. Thus, 1957-1976 represents the best-recorded set of documents, with the widest area coverage.

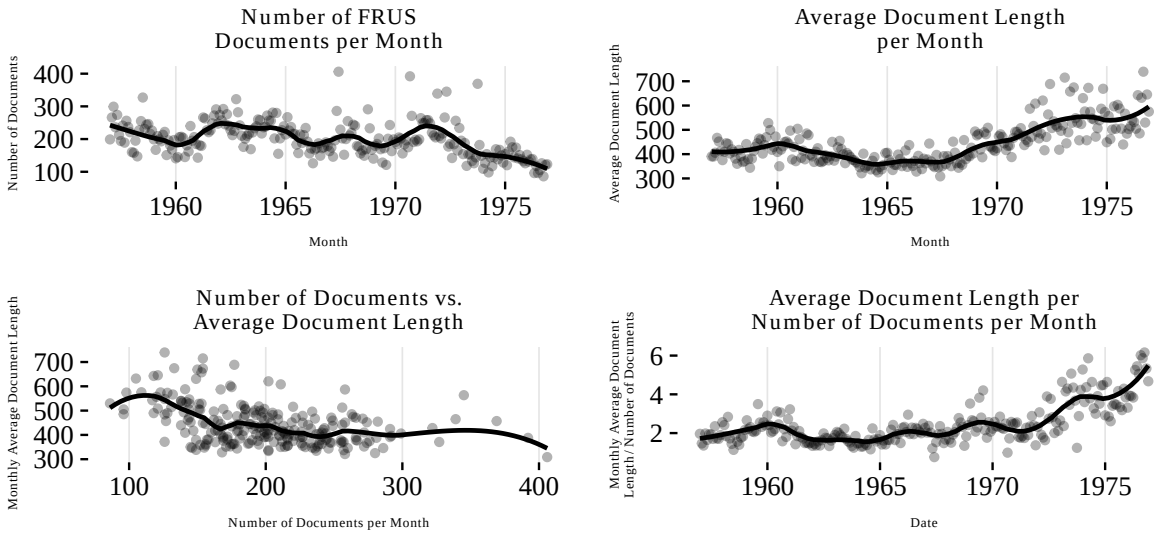

Figure 4.1: Monthly counts of documents, average length by month, and length vs. document counts.

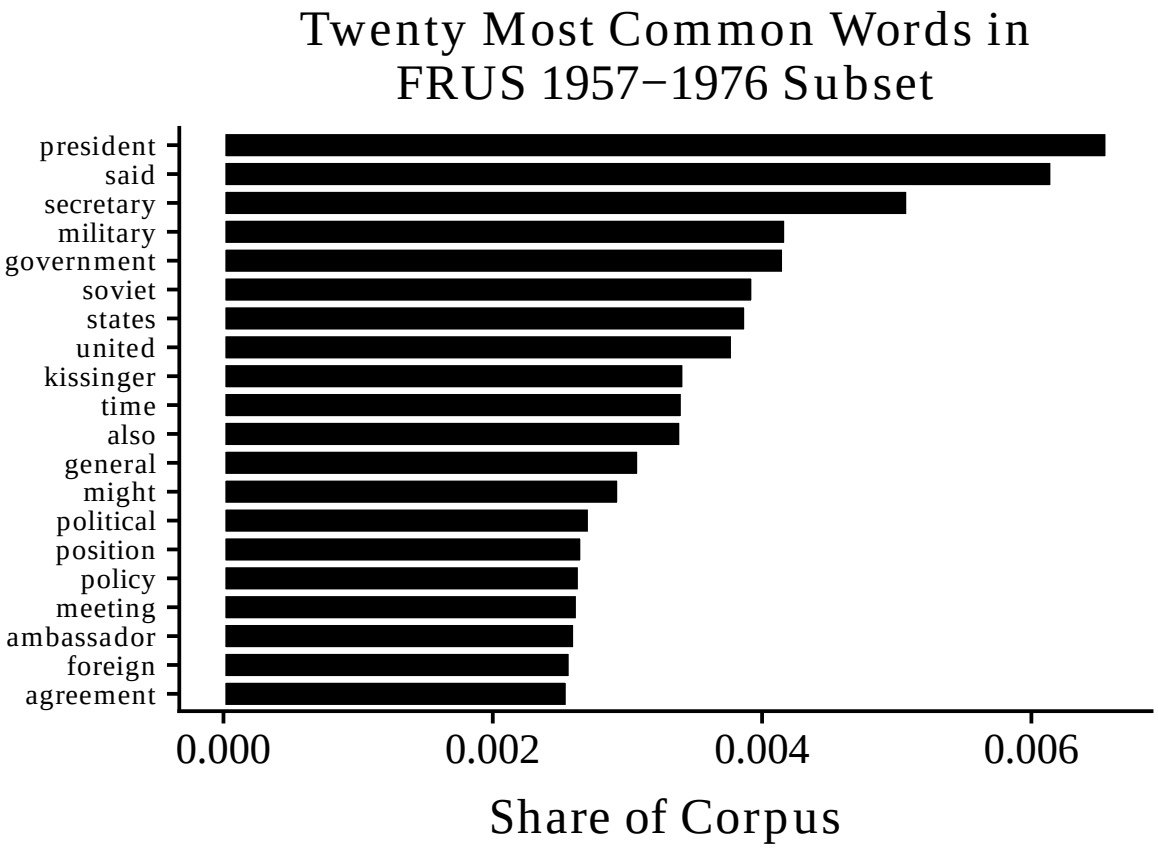

Figure 4.2: Most frequent words in 1957-1976 subset of FRUS corpus.

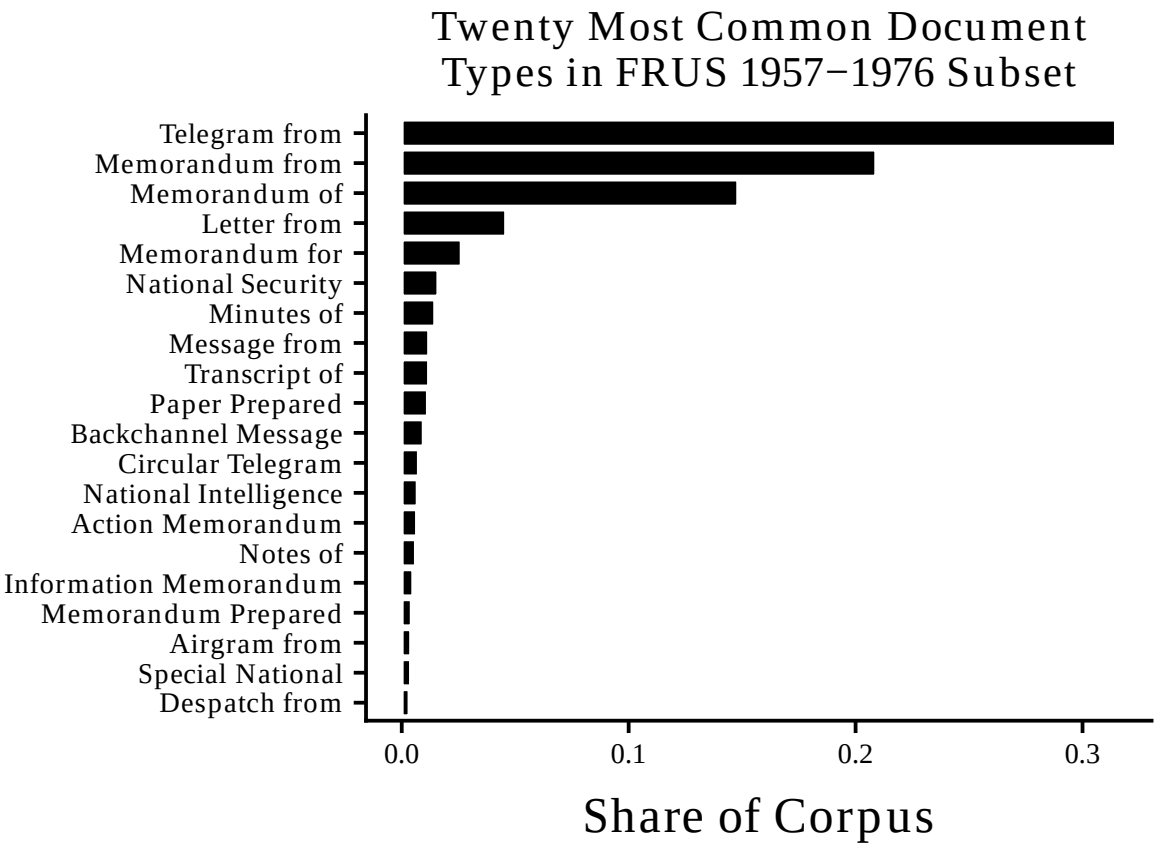

Figure 4.3: Most frequent document types in 1957-1976 subset of FRUS corpus.

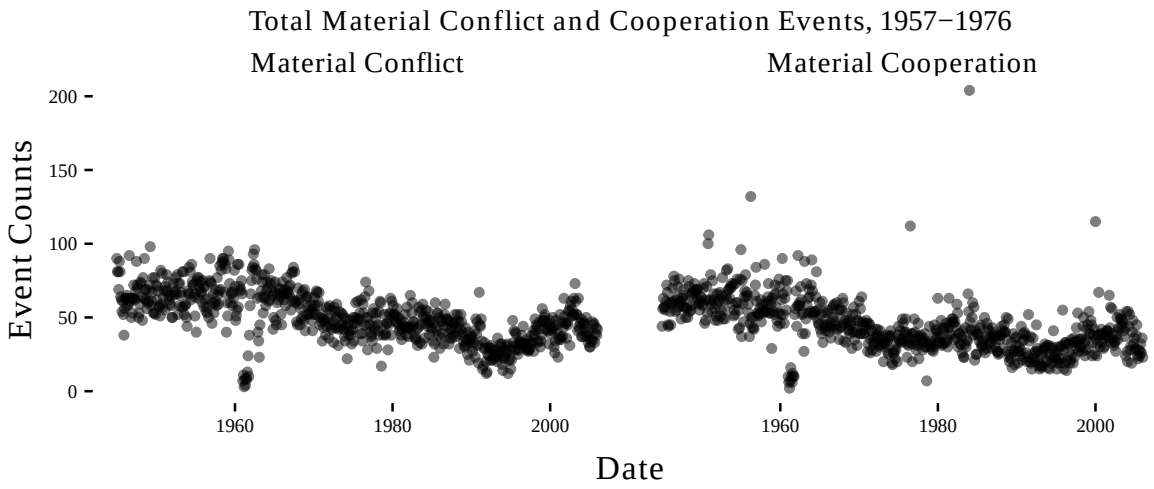

Figure 4.4: Total number of material conflict and material cooperation events per month, 1957-1976.

data consist of directed dyads, in the format *(DATE, STATE A, ACTION, STATE B)*, where *STATE A* is the sender, *STATE B* is the receiver, and action is one of four categories: material cooperation, material conflict, verbal cooperation, or verbal conflict.[8] For this analysis, I focus on material conflict and material cooperation, as these are less frequent than their verbal equivalents, and presumably, more costly. I first plot the number of global material cooperation and conflict events in Figure 4.4, they are similar, suggesting are measuring actions of similar costliness. I then examine the material conflict and material cooperation event counts for each of the three cases of interest Cuba, France, and China, in Figure 4.5. We see that there is consistent low-level conflict and cooperation from the United States, directed at each of these three countries.

[8]There is a fifth category, "made a statement," which is considered a neutral category. I ignore it in this analysis, as neutral statements are not particularly costly, nor could they indicate a shift in policy.

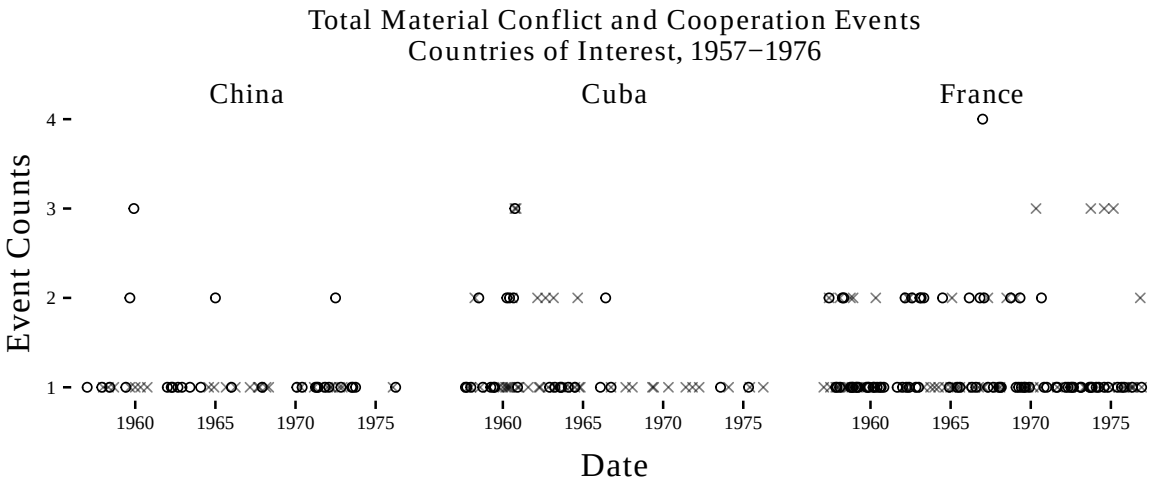

Figure 4.5: Total number of non-zero material conflict and material cooperation events per month, 1957-1976. X's are conflict counts, O's are Cooperation counts.

4.5.3 Combining Embeddings

The embeddings I use were initially estimated in three parts, 1957-63, 1964-69, and 1970-76, due to computational limits. To combine these embeddings, I turned to the Procrustes transformation, following work by Hamilton, Leskovec, and Jurafsky (2016) on combining embeddings to model language change. The goal of the Procrustes transformation is to rotate and scale one matrix to be as close to another as possible (Borg and Groenen, 2003, ch. 20). In this case, I align the 1964-69 and 1970-76 embedding matrices to the 1957-63 embedding matrix. I do this by subsetting all three embedding matrices to their shared words and use the Procrustes method to align these shared words. I then take the weights matrix used to transform the shared words matrix, and multiply the complete embedding matrix by these shared weights. Obviously, a concern in this case is that the Procrustes transformation distorts the underlying embeddings. To test this concern, I calculate the correlation

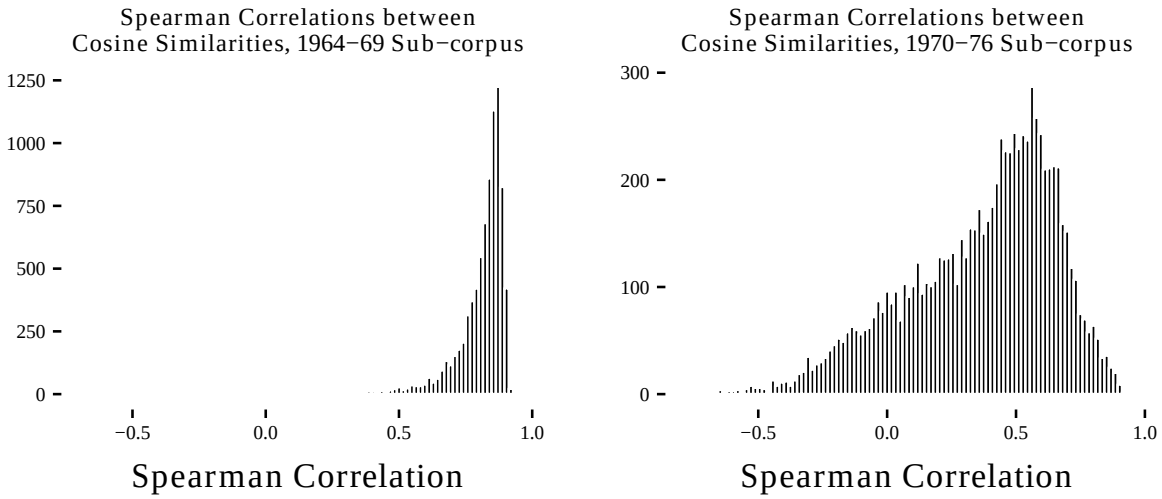

Figure 4.6: Spearman's rho, correlation between each word's cosine similarities between the original word and the procrustes-transformed corpus.

between the cosine similarities for the 1964-69 and 1970-76 sub-corpora. Examining Figure 4.6, we see that the the cosine similarities are more positively correlated for the 1964-69 subcorpus than the 1970-76 subcorpus. The least correlated similarities across both datasets are proper nouns like "Kissinger," reflecting the changing role of Henry Kissinger from the Eisenhower/Kennedy administrations to the Nixon/Ford administrations. More generally, this reflects the way in which language changes over time based on social factors, even over this small time period, and as such, choosing anchoring points matters.[9]

[9]One possible sensitivity analysis could examine alternate strategies for aligning matrices: which matrices are transformed, are transformations propagated, and, are certain kinds of words more likely to change over time?

4.6 Model

Measuring securitization is difficult, because existing models for modeling language, such as the structural topic model (Roberts, Stewart, and Airoldi, 2016), cannot measure the meanings of individual words. Many models treat words as atomic units, all equally (dis)similar, "democracy" is as similar to "autocracy" as it is to "kitten." Even models which can model latent themes at the document level (like topic models) do not allow for measurement of word-level similarity. Word-level embedding models offer an alternative which can capture the meanings of individual words.

Word-level models work by modeling the co-occurrence of pairs of words; they follow the "distributional hypothesis." The distributional hypothesis states that words with similar meanings will appear in similar contexts (Harris, 1954). Traditionally, this was accomplished by calculating the pointwise mutual information (PMI) of two words w_i, w_j: the log probability that two words co-occur within some span, divided by the probability of each of their occurrences: $PMI(w_i, w_j) = log\left(\frac{p(w_i,w_j)}{p(w_i)p(w_j)}\right)$ (Church and Hanks, 1990). PMI can be combined with matrix factorization techniques, such as singular value decomposition (SVD), which can capture relationships between words which never directly co-occur with each other, but only appear in similar contexts (Deerwester et al., 1990; Landauer and Dumais, 1997).

This factorization-based approach was the standard practice until recently (see Turney and Pantel (2010) for a review), when neural network methods became dominant (Mikolov, Yih, and Zweig, 2013; Mikolov et al., 2013). While neural language models had existed previously (Bengio et al., 2003; Collobert et al., 2011), they were slow, and did not scale to large corpora. This is because, when computing the probability that two words would co-occur, traditional neural networks calculated

co-occurrence as a softmax (multinomial) probability, comparing the probability of w_i and w_j co-occurring to the probability of w_i co-occurring with all other words in a corpus. Mikolov et al. (2013) introduced a variety of techniques in Word2vec to scale neural word embedding models, most prominently, negative sampling, which replaces the softmax function with an approximation. Negative sampling uses noise-contrastive estimation (Gutmann and Hyvärinen, 2010) to train a logistic regression to distinguish between positive cases (where w_i and w_j co-occur) and negative cases (where w_i and w_j do not co-occur), and downsamples the negative cases, speeding up inference, and improving accuracy.

The embeddings learned by Word2vec are coefficients, which can capture meaning the meanings of words, in relation to other words. This meaning is measured using the cosine similarity of two vectors:

$$cos(\boldsymbol{w}_i, \boldsymbol{w}_j) = \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{||\boldsymbol{w}_i||\ ||\boldsymbol{w}_j||}$$

which returns a number $[-1, 1]$, where 1 is most similar, -1 is most dissimilar. Not only can these vectors be compared directly, they are additive, and this additivity can capture relationships like "king $-$ man $+$ woman $\approx$ queen." These embeddings, then, can capture relationships in language which appeal to social scientists, such as emotion and bias (Garg et al., 2018; Rheault et al., 2016).

However, word embeddings, since they are multidimensional latent variable models, are not identified by default, and thus, unusable in the regression modeling framework favored by social scientists. This is a known problem with this class of model (Rivers, 2003). Embeddings are not identified for two reasons: an infinite combination of embedding values can result in the same observed data, and the vectors

from the model are not directly interpretable, the embeddings have neither a meaningfule direction nor scale. Embeddings cannot be directly used in a regression, as a regression coefficient would have no meaning.

4.6.1 Bayesian Word Embeddings

To measure securitization, I turn to Bayesian Word Embeddings, and the associated identification strategy introduced by Lauretig (2019). Bayesian Word Embeddings (BWE) offer several improvements over standard word embeddings for social scientists, including uncertainty estimates for word embeddings, priors to determine the number of appropriate dimensions, and a strategy to identify the resulting embeddings.

While there have been previous Bayesian embedding models, these have not explicitly modeled the uncertainty in embeddings, they have simply offered a better point estimate (Barkan, 2017; Havrylov and Titov, 2018; Ji et al., 2017; Rudolph et al., 2016). While Han et al. (2018) take inference and uncertainty seriously, they do not grapple with the identification problems associated with using embeddings in a regression setting.[10]

Bayesian Word Embeddings (Lauretig, 2019) cast word embeddings as a Bayesian latent variable model, in particular, as a model similar to a multidimensional ideal

[10]The approach proposed by Han et al. (2018) is also implemented in Python/Pytorch, which limits its utility for social scientists, many of whom have a workflow that is overwhelmingly (if not exclusively) in R.

point model, which is then estimated with variational inference. The likelihood for the model is

$$
\begin{aligned}
p(\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\alpha}_{\mathbf{X}},\boldsymbol{\alpha}_{\beta}|\mathbf{Y}) \propto \\
&(\mathbf{1}[z_{ij}>0]\mathbf{1}[y_{ij}=1]+\mathbf{1}[z_{ij}<0]\mathbf{1}[y_{ij}=0]) \\
&\mathcal{TN}(z_{ij}|\boldsymbol{x}_i^{\top}\boldsymbol{\beta}_j,1)\times \\
&\prod_i \mathcal{MVN}(\boldsymbol{x}_i|0,\boldsymbol{\alpha}_X^{-1})\times \\
&\prod_i \mathcal{MVN}(\boldsymbol{\beta}_j|0,\boldsymbol{\alpha}_{\beta}^{-1})\times \\
&\prod_k \text{Gam}(\alpha_{X_k}|c_{X_0},d_{X_0})\times \\
&\prod_k \text{Gam}(\alpha_{\beta_k}|c_{\beta_0},d_{\beta_0}).
\end{aligned}
\tag{4.1}
$$

In this model, $\boldsymbol{Y}$ is the observed data, and $c_{X_0}, d_{X_0}, c_{\beta_0}, d_{\beta_0}$ are the user-specified hyperparameters, which control the amount of variance explained by each dimension, an approach known as automatic relevance determination (Bishop, 1999; MacKay and Neal, 1994). Everything else is a latent variable, learned from the data. z_{ij}, the linear predictor for co-occurrence, is passed through a probit link function to estimate the probability that two words co-occur.[11]

Bayesian word embeddings are estimated with variational expectation-maximization. The goal of Bayesian inference is to solve the following integral

$$
P(\boldsymbol{Y}) = \int p(\boldsymbol{Y},\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{4.2}
$$

[11]While the logit link function is often the conventional choice, in this case, the probit is conditionally conjugate with the embeddings, which makes the math more tractable. Using a logit link would require an approximation, such as the one proposed by Jaakkola and Jordan (1997).

where $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_\beta\}$. This integral is analytically intractable. However, we can approximate a lower bound, using Jensen's inequality:

$$\begin{aligned} \ln P(\boldsymbol{Y}) &= \ln \int p(\boldsymbol{Y}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \ln \int \mathcal{Q}(\boldsymbol{\theta}) \frac{P(\boldsymbol{Y}, \boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &\geq \int \mathcal{Q}(\boldsymbol{\theta}) \ln \frac{P(\boldsymbol{Y}, \boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \mathcal{L}(Q) \end{aligned} \tag{4.3}$$

where $\mathcal{L}(Q)$ is evidence lower bound (ELBO).

The difference between the true model $P(Y)$ and variational approximation $\mathcal{L}(Q)$ is the Kullback-Leibler divergence:

$$KL(Q||P) = -\int \mathcal{Q}(\boldsymbol{\theta}) \ln \frac{P(Y|\boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{4.4}$$

which can be estimated using mean-field variational Bayes.[12]

4.6.2 Identifying Embeddings

Identifying Bayesian word embeddings is straightforward, drawing on work identifying multidimensional ideal point models (Clinton, Jackman, and Rivers, 2004; Rivers, 2003). We need to fix $K(K+1)$ points in the embeddings in order for the embeddings themselves to be identified: rotated and scaled so that the dimensions are linearly interpretable. Choosing anchors is difficult. However, when K, the number of dimensions, increases to the dimensionality common in many word embeddings. Lauretig (2019) proposes that theory is used to motivate initial anchor selection, but then,

[12] See Lauretig (2019) for the full MFVI factorization.

each remaining anchor is the most cosine dissimilar word to the previous word. Furthermore, Lauretig (2019) offers a way to create multiple anchors for each dimension, for example, scaling a dimension so that one end is "war" and the other is "peace." This allows individual dimensions of word embeddings to be used in regressions, and the regression coefficients will be interpretable.

4.7 Model Results

I fit the BWE model on each of the three subcorpora, with 50 embedding dimensions, a window of 5, and establish convergence when change in the ELBO drops below .0001. I then combine the resulting embeddings using the Procrustes results mentioned above. I anchor the embeddings on "conflict" and "cooperation", so positive numbers are more conflictual, and negative numbers more cooperative, and label this scaling "bellicosity." I then multiply the scaled embedding matrix by the binarized document-term matrix, scaling each document according to its bellicosity.

I begin by estimating the bellicosity over time, for all documents in the corpus, allowing the slopes to vary by president. I use a generalized additive model, implemented in `mgcv`, to relax the assumption of a linear time trend in bellicosity. I plot the predicted bellicosity for each president in Figure 4.7, along with 95% confidence intervals. The patterns displayed in the plot align with historical understanding of presidential policies, for example, Kennedy was initially more hawkish on foreign policy than Eisenhower, and as Nixon's term advanced, he opened China, and ended the Vietnam War. Furthermore, we see the largest changes in bellicosity occur when the president's party changes: Eisenhower to Kennedy, and then, Johnson to Nixon.

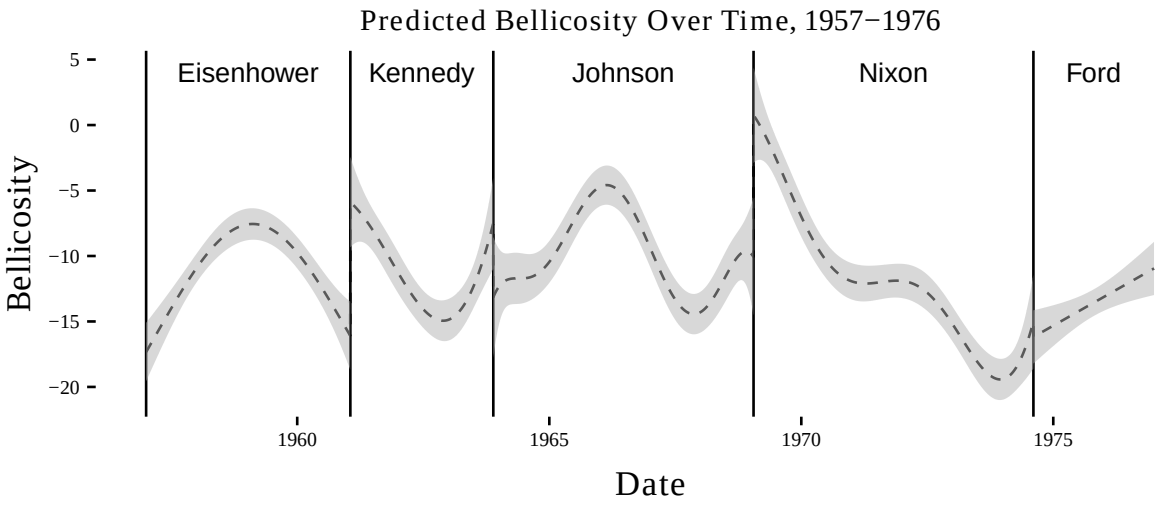

Figure 4.7: Predicted bellicosity from a generalized additive model for each president, 1957-1976, across the entire FRUS corpus.

4.7.1 Does Securitization Affect Action?

I begin by investigating the effect of securitization on policy choices: does securitization have material effects on policy? In particular, when relations with a country become more securitized, does resulting behavior become more conflictual or cooperative? There is evidence that language can predict action, Zimmerman and Palmer (1983) showed that CIA estimates of Soviet defense spending could be predicted from Soviet leaders' speeches. Unlike Zimmerman and Palmer (1983), who hand-coded the bellicosity in documents, I measure securitization using the bellicosity measures discussed previously. I examine behavior at the monthly level, where the outcome is counts of material and conflictual events as coded in the Cline Center event data. Events are directed dyads with the United States as the source, and the target country in each of the three cases. As in the previous section, I use generalized additive

models (Wood, 2006) here, with a negative binomial link function to model the relationship between bellicosity and action, and plot the predicted responses in Figure 4.8.[13]

We see a clear null effect, that bellicosity has no effect on either cooperation or conflict during the period under study, regardless of the effect of ideology or material conditions on securitizing rhetoric. What explains this null effect, across all three cases? I propose four explanations: that our qualitative understanding of securitization does not align with the quantitative measure, that I am using an incorrect dependent variable, that policy discussions do not matter, or, that the effect of securitization is not easily measured by the standard dependent variables political scientists examine.

4.7.2 Is There Securitization?

One possibility is that there is no securitization in the cases I have selected. This would be a disconcerting possibility, given the historical record. However, it is possible that these events were only securitized in retrospect. I test the presence of securitization by asking: when there is a securitizing event, does rhetoric become more bellicose? Investigating the three cases above, I subset the FRUS corpus to relevant documents for each country, using names, demonyms, and leaders' names to select these documents with regular expressions. I then include indicator variables for the relevant events: the Cuban Revolution, the French nuclear bomb, and the Chinese nuclear bomb. I estimate generalized additive models (Wood, 2006), where I regress

[13]Due to the relatively low number of events in a given month, I also dichotomized the outcome, coding a month as 1 if it contained any conflictual (cooperative) events, and 0 otherwise. I then model this outcome with a generalized additive model with a logit link function. The substantive results remained the same. Additionally, I fit these models with a zero-inflated Poisson link function, and the results did not change.

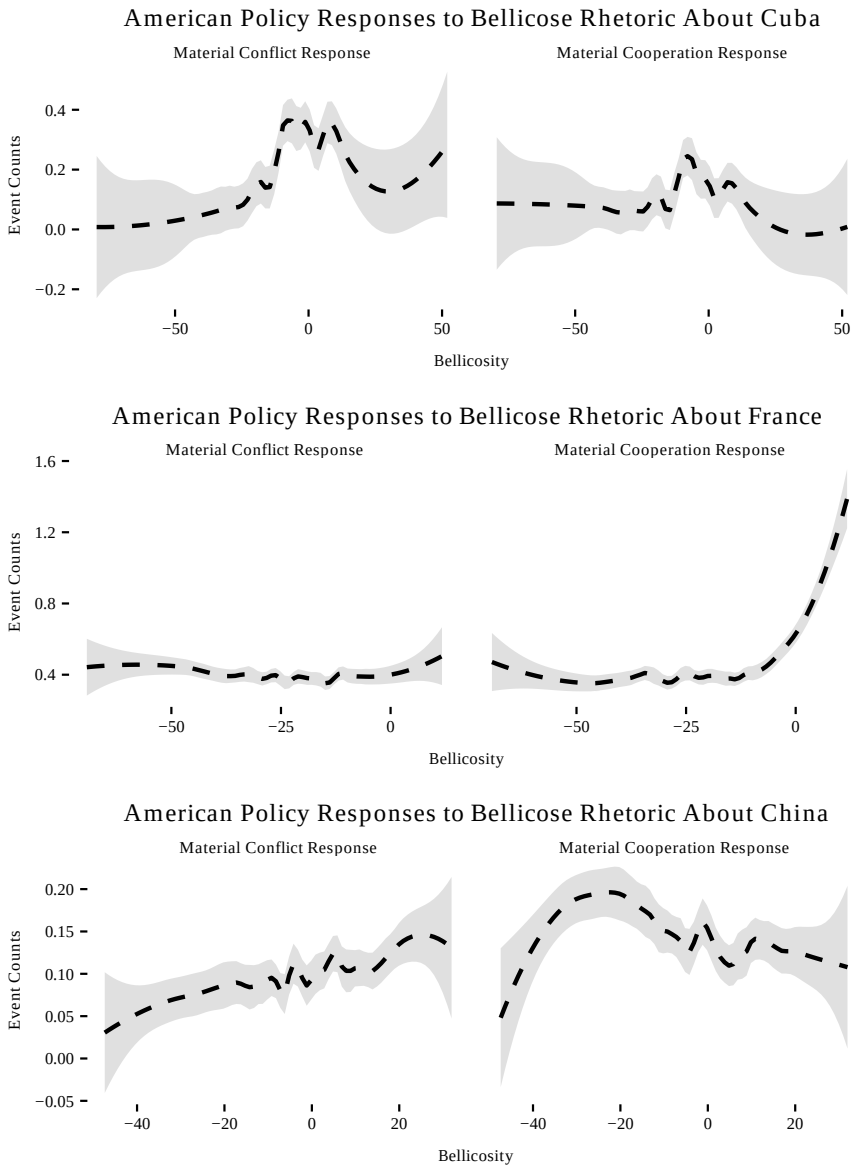

Figure 4.8: The null effect of bellicose rhetoric on conflictual and cooperative behavior across all three cases of interest. The y-axis is the number of predicted events from the model.

bellicosity on the interaction between this indicator and a nonlinear time-trend. In the case of France, I include an additional indicator for the war in Algeria (which ended in 1962), as this could influence discussions of bellicosity.

We see securitization in the ideological and ideological-material cases but not in the strictly material case. This contributes to the understanding of securitization as an idea-driven response to theories that root all behavior in material capabilities. In particular, across the two nuclearizing cases, it is China that sees an increase in bellicosity, given that the United States viewed it as another extension of Soviet plots to dominate the world. However, it is noteworthy that the effect of Chinese nuclearization is smaller than either the increase in bellicosity directed at Cuba, or the decrease directed at France.

What explains the decrease in bellicosity directed at France? One possibility is that rhetoric became less bellicose after nuclearization, when it was not possible to change the fact of French nuclearization. That is, once France became a nuclear power, it no longer made sense to contest the issue.

A second possibility is that nuclearization is affected by the close relationship between the United States and France. In the cases in the time period covered, the only cases of nuclearization occur in either close American allies (UK, France, Israel), or in avowed enemies (USSR, China), which makes isolating the effects of nuclearization alone difficult. The ideal test case to understand the effect of nuclearization on securitization would be a neutral country, without strong ideological commitments to either side. Unfortunately, such a case does not exist. When we look at the effect of nuclearization in China, we see a small statistically significant effect, suggesting

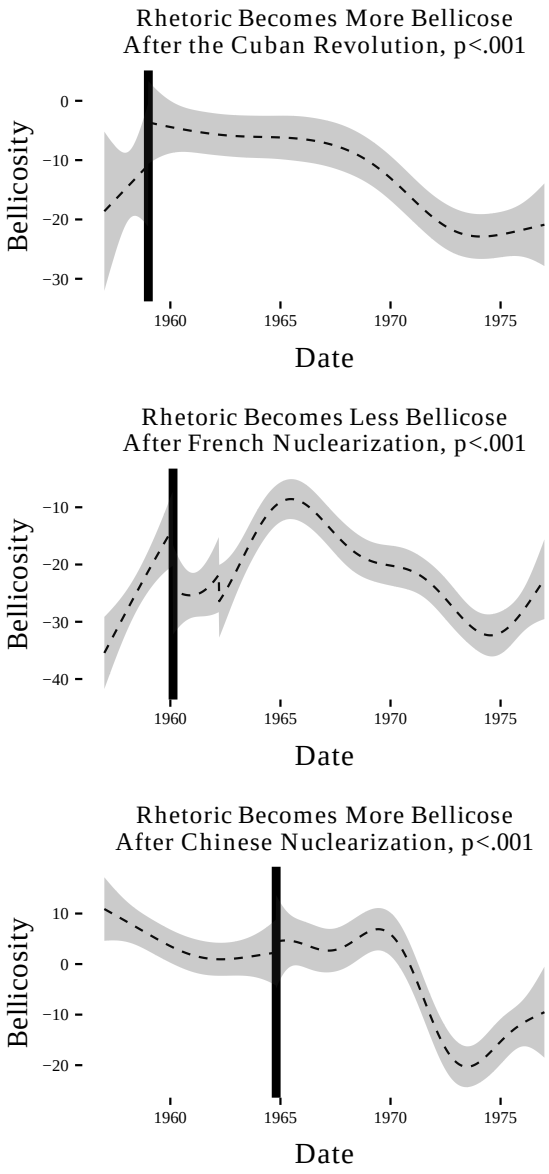

Figure 4.9: In ideological cases (China and Cuba), there is an increase in bellicose rhetoric, suggesting that ideology contributes to securitization. In material-only cases, there is a decrease in bellicose rhetoric.

that while nuclearization results in a small increase in bellicosity when the nuclearizing power was an ideological foe, ideology seems be the reason for the difference in amount of securitization between China and France.

More generally, this highlights what Buzan (2008) observed: that seemingly materialist relationships are, in fact the product of ideological relationships. This highlights the importance of understanding ideas in international relations, that material capabilities are not the end-all, be-all of relationships in international politics.

4.7.3 Alternate Dependent Variable

A second possibility is measurement error in the either the dependent or independent variables, in particular that the study focuses on the wrong dependent variable. To test this assumption, I examine the effect of securitizing rhetoric on the probability of MID onset between the United States and each of the three countries under consideration. MIDs are a common dependent variable in quantitative studies of conflict, and encompass everything from verbal threats to war, defined as $>$ 1000 battlefield deaths (Singer, 1972). MIDs differ from the Cline Center event data by focusing on larger "macro" events (such as war). While there may be fewer overall events recorded in the dataset, then, they are more noticeable events, and thus, if bellicosity leads to action, these larger actions are more likely to be the result of a change in securitizing rhetoric. Additionally, as MIDs and Cline Center events are coded according to different criteria, they are unlikely to measure the exact same phenomena, while still capturing the same underlying concepts, helping to confirm the validity of the results. I take an expansive definition of MID onset, and code onset as a transition from 0 to any level of MID. Here, I use a generalized additive

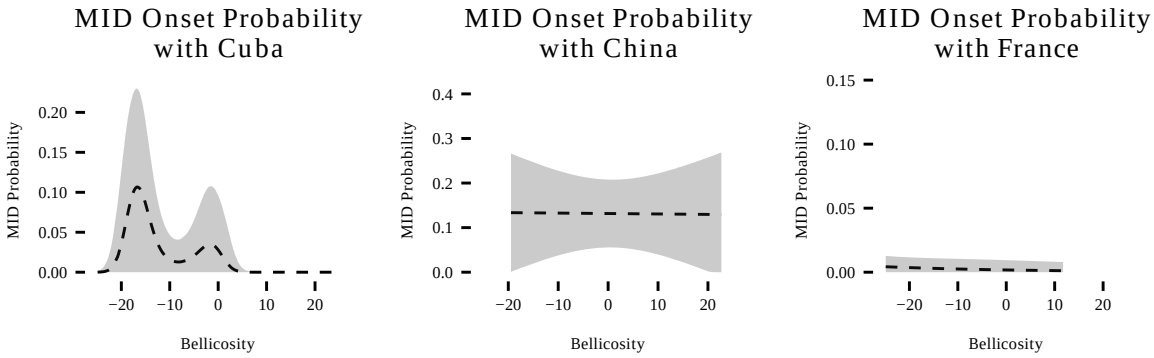

Figure 4.10: The null effect of bellicose rhetoric on the probability of MID onset across all three cases of interest.

model with a logistic link function to estimate the effect of bellicose rhetoric on the probability of MID onset, and I display the relationship in Figure 4.10.

We see that securitization again has no effect on the probability of MID onset across all three cases. This reassures us that the dependent variable is not the reason for this null response, at least in the case of conflict. Unfortunately, there is no equivalent of the MIDs data for cooperation. However, based on the similarity of results for Cline Center events and MIDs, it seems unlikely that results would be different.

4.7.4 Do Discussions Matter?

The third possibility is that policy discussions, or (more generously) policy discussions in FRUS, do not matter for understanding state behavior. This is possible, that much of the discussion in the FRUS corpus is not relevant. Chaney et al. (2016) note that in the full 1975-79 corpus they use, that many of the documents are irrelevant. It is possible that even with the expert document selection in the FRUS corpus, there are still many documents which are not necessarily relevant. The difficulty here come

from defining relevance. If we define relevant documents as "those that affect the outcome," we have defined our model in terms of itself, which would invalidate our inferences.

4.7.5 Securitization and Power

The fourth possible reason for the null effect is that securitization matters, but not in a way that can be measured using conventional international relations metrics. Understanding this requires drawing on *power theory*, in particular the *Three Faces of Power* (Lukes, 2004), and the extension by Barnett and Duvall (2005). Standard models of power in quantitative international relations focus on the first face of power, a behavioral approach, where one actor compels another to do something they otherwise would not. Modeling responses to securitization as material events (conflict and cooperation) largely falls into this first face of power category. The assumption is that an issue becomes securitized, and the United States then uses force to change the other actor's behavior, and therefore securitization is translated into direct action.

However, this is not the only way securitization can shape policy. Securitization can also shape the decision-making of actors, and lead them to make decisions which go against their own interest, what Lukes (2004) calls the "third face of power." Here, securitization can lead to actors making decisions which appear irrational when considered from a standard materialist perspective. For example, the securitization of Cuba after the 1959 Revolution led to a 50-year period where any effort to open relations with Cuba was unthinkable, until 2014 (Baker, 2014), despite the Cold War ending more than a decade previously. Securitization, then, can tie the hands of future actors, even if they were not party to the original securitizing discourse. Similarly,

when an ideological foe like China develops nuclear weapons, this can lead to a lack of engagement. After all, Nixon opening relations with China was seen as exceptional, and only something a hardened Cold Warrior could pull off (Herring, 2008, ch. 17).

This limitation raises fascinating questions for the integration of securitization theory into quantitative international relations. Traditionally, international relations has focused on observed actions, and the questions of what motivate them. However, if securitizing an issue changes the mental map of policymakers to the extent that they are unwilling to consider certain policy options when responding to an issue, this suggests that existing models of international conflict, which generally takes a behavioral tack towards understanding behavior, are insufficient for understanding foreign policy decision-making.

More generally, this raises questions about what modeling securitization looks like in a more general case, especially when handling a case where there is not a discrete event leading to the beginning of the securitization process. If the effect of securitization is to constrain decision-makers, then we are forced to think about the non-events in our datasets: are they random (they could have happened), or are they non-events because policy-makers have foreclosed options available to them? The difficulty of isolating this distinction suggests that in the future, scholars studying foreign policy decision-making would find it useful to access the text of decision-makers' deliberations, and model decision-making processes such as securitization, rather than relying on crude country-year level indicators.

4.8 Conclusion

In this paper, I set out to measure the role of securitization in foreign policy decision-making. I drew on the documents collected in the *Foreign Relations of the United States*, a collection of primary source documents about American foreign policy, the Cline center event data, and a Bayesian Word Embedding model developed by Lauretig (2019). I examined a period from 1957-1976, where I trained embeddings on three subsets of the FRUS corpus, and then aligned these embeddings using the Procrustes method. I identified and scaled the embeddings using the method discussed in Lauretig (2019). I then investigated three cases: the 1959 Cuban Revolution, France's acquisition of the nuclear bomb, and China's acquisition of the nuclear bomb, investigating cases with variation in either ideology, material capability, or both. I tested whether securitization led to a change in American foreign policy behavior.

However, in none of the three cases do we see an effect of securitization on the *behavior* of states. This finding holds for both conflict and cooperation, and even when conflict is measured across multiple dependent variables, the null finding holds. This null finding is not simply due to an absence of securitization among decision-makers: there was an increase in securitization when ideology was salient, but not when the the change was simply due to a change in material capabilities. I suggest this is because securitization constrains behavior in ways that cannot be directly measured. For example, the securitization of Cuba after the Revolution prevented any rapprochement between the two countries; domestic pressure in the United States was too strong. This ties back to a key insight of securitization theory (Buzan and Herring, 1998), that much of what looks like behavior shaped by material capabilities,

is often shaped by securitization, and not accounting for this can lead us to wrong inferences.

More generally, this work sets up exciting avenues for future quantitative research on securitization. Two possible avenues for this research include extending the models and data to broader swathes of American history, and considering securitization when there is not a single triggering event.

Extending the time frame under study is one appealing idea, in order to observe cycles of isolationism and engagement in American foreign policy. However, the current limitation for extending the timeframe under consideration is data availability: FRUS documents are only available consistently until 1976, with spottier coverage for the end of the Cold War, and no post-Cold War coverage. Unfortunately, this means that any attempt to understand securitization from 1945-1989 is confounded by the presence of Cold War tensions. In the pre-Cold War era, while the FRUS has been digitized, the only consistently available dependent variable is MID onset, and unfortunately, MID onset is a rare variable, and it does not capture cooperation.

Investigating securitization when there is not a single triggering event is another promising vein of research. Securitization theory has been applied to questions of climate change and migration, topics without a clear triggering event to begin the securitizing process. In these cases, understanding what securitization looks like could provide valuable insight into how policymakers securitize non-security issues.

Chapter 5: Conclusion

In this dissertation, I set out to understand the role of elite decision-making in American foreign policy. I developed a novel statistical tool, Bayesian Word Embeddings, and used it, alongside other tools, to elite decision-making in American foreign policy. Using the recently digitized *Foreign Relations of the United States*, I examined both the practice of securitization, and elite casualty sensitivity during the Vietnam War, explicitly focusing on the text of policymakers' decisions. I found that securitization has no effect on either cooperation or conflict by the United States, even in cases where the relations between the United States and another country are securitized due to ideology. I discuss possible reasons for this result, and what it means for how we model decision-making in American foreign policy. I also found that while casualties affect bombing decisions during the Vietnam War, casualties have no effect on elite discussions about ending the war.

This complicates our understanding of elite decision-making in American foreign policy. Traditionally, elites have been treated as a black box (Gartner, 1998): scholars can observe inputs (battlefield outcomes, public opinion, material capabilities), and outcomes (war termination), and infer the role of elites. However, my findings show that outcomes and their inputs are not directly related: the way many quantitative scholars model foreign policy is not just an approximation of reality, but a distortion.

Furthermore, these findings have surprising implications for qualitative scholars who study American foreign policy (Jervis, 1976; Saunders, 2011; Yarhi-Milo, 2014). Many of these scholars highlight selected documents and speeches and tie them to foreign policy actions. However, by ignoring the broader context for these documents and actions, they may overstate an effect. Using text as data tools and a quantitative approach allows us to not only model all the documents, we can also measure the magnitude of a policy response. Furthermore, in quantitative work, we can also measure all the cases where there is no policy response: in the securitization case, the months where there are no material cooperation or conflict events. Ignoring these non-events can lead us to overvalue the relationship between policy discussions and observed outcomes.

Future work would continue this line of research into elite decision-making, taking into account the historical breadth of the FRUS corpus, investigating the broader elite-public relationship, elite responses to casualties, and how these have changed over time. As LaFeber (2013) notes, the United States was not isolationist before World War I. However, its foreign policy, was motivated explicitly by business interests, rather than the appeals to liberalism that characterized post-1945 foreign policy, and it could be interesting to quantitatively test hypotheses about the 19^{th} century, a period that has received a great deal of qualitative interest (Buzan and Lawson, 2013), but far less quantitative interest.

A second line of future work would exploit the structure of the communication network in FRUS. The metadata for the directed FRUS documents (those which are "from" and "to" somebody) could be extracted, allowing us to not only model what is said, but to whom. The research on elite deliberations, the importance of advisers,

and groupthink in decision-making (Allison and Zelikow, 1999; Jervis, 1976; Saunders, 2017), modeling both the text and network structure of communication could enable scholars to measure how communication changes as groupthink sets in (as in the case of the Vietnam War) and, how much policy can change due to groupthink or changes in communication patterns among decision-makers.

However, these empirical applications would require extending the Bayesian Word Embedding model. The existing model offers a way to identify word embeddings, and use the identified embeddings to conduct inference, which makes embeddings useful to social scientists in ways that the structural topic model made topic models useful for social scientists (Roberts, Stewart, and Airoldi, 2016). Furthermore, by providing an `R` package, I offer a public good for political scientists that is neither proprietary and unavailable (Barkan, 2017), nor written in Python (Han et al., 2018; Rudolph et al., 2016).

However, there are several ways in which the existing model could be extended. These include scaling model estimation, changing the underlying distributions, incorporating word and document metadata into the embeddings, and developing a causal inference framework for word embeddings.

The existing BWE software is relatively slow. This is because the current variational inference implementation requires estimating every parameter for every iteration, which is incredibly slow, especially as corpora grow larger. For example, training each of the 6-year slices of the FRUS corpus in Chapter 4 took $\approx$ 1 week on Ohio State's *Unity* cluster, after extensive preprocessing to reduce their size. This approach is inconvenient for social scientists without access to supercomputing resources. One

way to reduce the amount of training time required to estimate the model is to estimate the parameters using stochastic variational inference (SVI) (Hoffman et al., 2013). Rather than updating every parameter at every iteration, SVI randomly sample a small number of words, updates them, and then computes a weighted average of the new and the old word estimates. SVI can be extended by incorporating an adaptive learning rate, learning the weights of the weighted average as the model is estimated, which can further decrease training time and improve model fit (Ranganath et al., 2013).

SVI is not the only improvement that can enhance model fit, we could also change the probability distribution underlying the model. Currently, the BWE model uses multivariate normal distributions for the embeddings, with a gamma Automatic Relevance Determination prior. While this choice is convenient, it also requires tuning to work well; if the user supplies too many dimensions, the model suffers from "rank collapse," and the rank of the embedding matrices is smaller than the dimensionality of the embedding. This can lead to issues when using the identification algorithm in Lauretig (2019). An alternate distribution which could be used to represent the embeddings is the von Mises-Fisher (vMF) distribution, a distribution on the unit (hyper)sphere. Each dimension of the embedding is orthogonal, and the model itself directly optimizes cosine similarity, the metric we use for embedding similarity, leading to improved performance (Banerjee et al., 2005) and more stable parameter estimates (Xu and Durrett, 2018). Unfortunately, estimating parameters from a vMF distribution is difficult, as the *concentration parameter* of the vMF distribution is not conditionally conjugate (it involves a Bessel function), and thus, an approximation is necessary (Gopal and Yang, 2014). What is unclear from existing work using the

vMF distribution for embeddings (Gopal and Yang, 2014; Xu and Durrett, 2018) is what inference for the vMF distribution requires, when vMF has been used in natural language processing models, it only generates point estimates, not posteriors.

While SVI and the vMF distributions would improve estimation, incorporating metadata into BWE would make it a "one stop shop" for social scientists interested in inference on language. Current approaches for incorporating document metadata into embeddings partition the embeddings based on the metadata, and are limited to categorical covariates (Han et al., 2018; Rudolph et al., 2017). Following the lead of Roberts, Stewart, and Airoldi (2016), the ideal method for including covariates would allow for both categorical and continuous covariates, as well as flexible specification of the relationship between covariates and embeddings. A strategy for estimating the embeddings could then share word embeddings across two tasks: predicting word co-occurrence, and then, predicting the probability that the word occurs in a document, using the metadata to inform the second estimate, while sharing the word embedding parameters across the two tasks. Estimating the effect of covariates on embeddings could prove difficult. However, as the embeddings are multidimensional, and document embeddings cannot necessarily be anchored and identified the same way word embedding are in Lauretig (2019). A post-estimation inference procedure to combine document and word embeddings would be the simplest solution. However, more work would be necessary to understand what the estimand of interest would be in this case.

A second way to incorporate metadata is at the word-level. This is less common in the social sciences. However, information like part-of-speech tags (Handler et al., 2016; Levy and Goldberg, 2014), language of origin, or simply distance between

word pairs (Cotterell et al., 2017) could provide valuable additional information to improve the quality of word embeddings. There are a variety of ways to incorporate this information: Levy and Goldberg (2014) measures distance on a parse tree, rather than simple co-occurrence, and Cotterell et al. (2017) use a tensor/multi-dimensional array structure, where each slice of the tensor is co-occurrence at a particular distance. Learning the embeddings is then treated as a tensor decomposition task. Here, it is not clear whether one would perform statistical inference with word-level metadata, or simply treat word-level information as "controls," information to be conditioned on to improve embeddings, whose effect would not be directly estimated in a post-estimation inference.

Why would we care about metadata and conducting inference with language? One reason is that we want to move beyond simple correlations, and want to make causal claims about what can be done with language. Once we understand how to estimate the effect of metadata on language, the second part of this project involves interrogating what causality and counterfactuals mean when discussing language. Currently, this is an open question, with existing work largely focused on topic-based causal models of language (Egami et al., 2018; Fong and Grimmer, 2016). However, there is room to expand on this work when the focus is on individual words.

Bibliography

About the Foreign Relations of the United States Series.

Chief of Army Praises Close Air Support Systems (1973).

Airoldi, Edoardo M. and Jonathan M. Bischof (2016). "Improving and Evaluating Topic Models and Other Models of Text". In: *Journal of the American Statistical Association* 111.516, pp. 1381–1403.

Aldrich, John H., Jacob M. Montgomery, and David B. Sparks (2014). "Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses". In: *Political Analysis* 22.4, pp. 435–456.

Aldrich, John H. et al. (2006). "Foreign policy and the electoral connection". In: *Annual Review of Political Science* 9, pp. 477–502.

Allison, Graham T. and Philip Zelikow (1999). *Essence of decision: Explaining the Cuban missile crisis*. Longman New York.

Althaus, Scott et al. (2017). *Cline Center Historical Phoenix Event Data*.

Andrews, Mark and Gabriella Vigliocco (2010). "The hidden Markov topic model: A probabilistic model of semantic representation". In: *Topics in Cognitive Science* 2.1, pp. 101–113.

Aradau, Claudia and Jef Huysmans (2014). "Critical methods in International Relations: The politics of techniques, devices and acts". In: *European Journal of International Relations* 20.3, pp. 596–619.

Austin, John Langshaw (1975). *How to do things with words*. Oxford university press.

Bafumi, Joseph et al. (2005). "Practical issues in implementing and understanding Bayesian ideal point estimation". In: *Political Analysis* 13.2, pp. 171–187.

Baker, Peter (2014). "US Will Restore Full Relations With Cuba, Erasing a Last Trace Of Cold War Hostility". In: *The New York Times*.

Balzacq, Thierry (2010). *Understanding Securitisation Theory: How Security Problems Emerge and Dissolve*. Routledge.

Banerjee, Arindam et al. (2005). "Clustering on the unit hypersphere using von Mises-Fisher distributions". In: *Journal of Machine Learning Research* 6.Sep, pp. 1345–1382.

Barkan, Oren (2017). "Bayesian Neural Word Embedding". In: *AAAI*, pp. 3135–3143.

Barkin, J. Samuel and Laura Sjoberg (2015). "Calculating critique: thinking outside the methods matching game". In: *Millennium* 43.3, pp. 852–871.

Barnett, Michael and Raymond Duvall (2005). "Power in international politics". In: *International organization* 59.1, pp. 39–75.

Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1, p. 38.

Bengio, Yoshua et al. (2003). "A neural probabilistic language model". In: *Journal of machine learning research* 3.Feb, pp. 1137–1155.

Benoit, Kenneth et al. (2018). "quanteda: An R package for the quantitative analysis of textual data". In: *Journal of Open Source Software* 3.30, p. 774.

Berinsky, Adam J. (2007). "Assuming the Costs of War: Events, Elites, and American Public Support for Military Conflict". In: *Journal of Politics* 69.4, pp. 975–997.

Biddle, Stephen and Stephen Long (2004). "Democracy and military effectiveness: A deeper look". In: *Journal of Conflict Resolution* 48.4, pp. 525–546.

Bishop, CM (1999). "Variational principal components". In: *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*. Vol. 1. IET, pp. 509–514.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877.

Blei, David M. and John D. Lafferty (2006a). "Correlated topic models". In: *Advances in neural information processing systems* 18, p. 147.

— (2006b). "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 113–120.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Borg, Ingwer and Patrick Groenen (2003). "Modern multidimensional scaling: Theory and applications". In: *Journal of Educational Measurement* 40.3, pp. 277–280.

Bražinskas, Arthur, Serhii Havrylov, and Ivan Titov (2017). "Embedding Words as Distributions with a Bayesian Skip-gram Model". In: *arXiv preprint arXiv:1711.11027*.

Brooks, Stephen G and William C Wohlforth (2001). "Power, globalization, and the end of the Cold War: Reevaluating a landmark case for ideas". In: *International Security* 25.3, pp. 5–53.

Bueno De Mesquita, Bruce et al. (1999). "An institutional explanation of the democratic peace". In: *American Political Science Review* 93.04, pp. 791–807.

Bumiller, Elisabeth (2009). "US lifts photo ban on military coffins". In: *The New York Times*.

Burr, William and Jeffrey T. Richelson (2001). "Whether to "strangle the baby in the cradle": the United States and the Chinese nuclear program, 1960–64". In: *International Security* 25.3, pp. 54–99.

Buzan, Barry (2004). *The United States and the great powers: world politics in the twenty-first century*. Polity.

— (2008). "The changing agenda of military security". In: *Globalization and Environmental Challenges*. Springer, pp. 553–560.

Buzan, Barry and Eric Herring (1998). *The arms dynamic in world politics*. Lynne Rienner Publishers.

Buzan, Barry and George Lawson (2013). "The Global Transformation: The Nineteenth Century and the Making of Modern International Relations 1". In: *International Studies Quarterly* 57.3, pp. 620–634.

Buzan, Barry, Ole Wæver, and Jaap De Wilde (1998). *Security: a new framework for analysis*. Lynne Rienner Publishers.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186.

Caverley, Jonathan D. (2014). *Democratic militarism: voting, wealth, and war*. Vol. 131. Cambridge University Press.

Caverley, Jonathan D. and Yanna Krupnikov (2017). "Aiming at Doves: Experimental Evidence of Military Images' Political Effects". In: *Journal of Conflict Resolution* 61.7, pp. 1482–1509.

Chaney, Allison et al. (2016). "Detecting and characterizing events". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1142–1152.

Church, Kenneth Ward and Patrick Hanks (1990). "Word association norms, mutual information, and lexicography". In: *Computational linguistics* 16.1, pp. 22–29.

Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). "The statistical analysis of roll call data". In: *American Political Science Review* 98.2, pp. 355–370.

Collobert, Ronan et al. (2011). "Natural language processing (almost) from scratch". In: *Journal of Machine Learning Research* 12.Aug, pp. 2493–2537.

Cotterell, Ryan et al. (2017). "Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis". In: *EACL 2017* 175.

Daddis, Gregory A. (2011). *No Sure Victory: Measuring US Army Effectiveness and Progress in the Vietnam War*. Oxford University Press.

Dafoe, Allan and Devin Caughey (2016). "Honor and war: Southern US presidents and the effects of concern for reputation". In: *World politics* 68.2, pp. 341–381.

Dafoe, Allan, Jonathan Renshon, and Paul Huth (2014). "Reputation and status as motives for war". In: *Annual Review of Political Science* 17, pp. 371–393.

Davenport, Tiffany C. (2015). "Policy-Induced Risk and Responsive Participation: The Effect of a Son's Conscription Risk on the Voting Behavior of His Parents". In: *American Journal of Political Science* 59.1, pp. 225–241.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, p. 391.

Defense Casualty Analysis System (2008). ARC Identifier: 2240992. National Archives & Department of Defense.

Dell, Melissa and Pablo Querubin (2017). "Nation building through foreign intervention: Evidence from discontinuities in military strategies". In: *The Quarterly Journal of Economics* 1, p. 64.

Deudney, Daniel (1990). "The case against linking environmental degradation and national security". In: *Millennium* 19.3, pp. 461–476.

DeYoung, Karen (2014). "Obama moves to normalize relations with Cuba as American is released by Havana". In: *Washington Post* 17.

DiCicco, Jonathan M. and Benjamin O. Fordham (2018). "The Things They Carried: Generational Effects of the Vietnam War on Elite Opinion". In: *International Studies Quarterly*.

Druckman, James N. and Lawrence R. Jacobs (2015). *Who Governs?: Presidents, Public Opinion, and Manipulation*. University of Chicago Press.

DuBrin, Doug (2008). "The Powell Doctrine. Background, Application and Critical Analysis". In: *News Hour Extra*.

Egami, Naoki et al. (2018). "How to Make Causal Inferences Using Texts". In: *arXiv preprint arXiv:1802.02163*.

Fearon, James D. (1995). "Rationalist explanations for war". In: *International organization* 49.03, pp. 379–414.

Feaver, Peter D. and Christopher F. Gelpi (2006). *Choosing your battles: American civil-military relations and the use of force*. Princeton University Press.

Fong, Christian and Justin Grimmer (2016). "Discovery of Treatments from Text Corpora." In: *ACL (1)*.

Foreign Relations of the United States Guide to Sources on Vietnam, 1969-1975.

Garg, Nikhil et al. (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Gartner, Scott Sigmund (1998). "Opening up the black box of war". In: *Journal of Conflict Resolution*, pp. 252–258.

Gartner, Scott Sigmund and Gary M. Segura (1998). "War, casualties, and public opinion". In: *Journal of Conflict Resolution* 42.3, pp. 278–300.

— (2000). "Race, casualties, and opinion in the Vietnam War". In: *Journal of Politics* 62.1, pp. 115–146.

Gelman, Andrew and Hal Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant". In: *The American Statistician* 60.4, pp. 328–331.

Gelpi, Christopher F. and Peter D. Feaver (2002). "Speak softly and carry a big stick? Veterans in the political elite and the American use of force". In: *American Political Science Review* 96.4, pp. 779–793.

Gelpi, Christopher F., Peter D. Feaver, and Jason Reifler (2006). "Success matters: Casualty sensitivity and the war in Iraq". In: *International Security* 30.3.

Gentile, Gian P. (2013). *Wrong Turn: America's Deadly Embrace of Counterinsurgency*. New Press, The.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy (2017). *Text as data*. Tech. rep. National Bureau of Economic Research.

Gerring, John (2017). "Qualitative methods". In: *Annual Review of Political Science* 20, pp. 15–36.

Gopal, Siddharth and Yiming Yang (2014). "Von mises-fisher clustering models". In: *International Conference on Machine Learning*, pp. 154–162.

Graeger, Nina (1996). "Environmental security?" In: *Journal of Peace Research* 33.1, pp. 109–116.

Grimmer, Justin (2010). "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases". In: *Political Analysis* 18.1, pp. 1–35.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". In: *Political Analysis*, mps028.

Gruber, Amit, Yair Weiss, and Michal Rosen-Zvi (2007). "Hidden Topic Markov Models". In: *AISTATS*. Vol. 2, pp. 163–170.

Gutmann, Michael and Aapo Hyvärinen (2010). "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304.

Haberman, Clyde (2017). *John Kelly Suggests More Americans Should Have the Honor of Serving. He's Right.*

Hafner-Burton, Emilie M. et al. (2014). "Decision maker preferences for international legal cooperation". In: *International Organization* 68.04, pp. 845–876.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 1489–1501.

Han, Rujun et al. (2018). "Conditional Word Embedding and Hypothesis Testing via Bayes-by-Backprop". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4890–4895.

Handler, Abram et al. (2016). "Bag of What? Simple Noun Phrase Extraction for Text Analysis". In: Association for Computational Linguistics.

Harris, Zellig S. (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162.

Haun, Phil and Colin Jackson (2016). "Breaker of Armies: Air Power in the Easter Offensive and the Myth of Linebacker I and II in the Vietnam War". In: *International Security* 40.3, pp. 139–178.

Havrylov, Serhii and Ivan Titov (2018). "Embedding Words as Distributions with a Bayesian Skip-gram Model". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1775–1789.

Herring, George C. (1990). ""Peoples Quite Apart": Americans, South Vietnamese, and the War in Vietnam". In: *Diplomatic History* 14.1, pp. 1–23.

— (2008). *From colony to superpower: US foreign relations since 1776*. Oxford University Press.

Higgins, Trumbull (1987). *The perfect failure: Kennedy, Eisenhower, and the CIA at the Bay of Pigs*. Norton New York.

Hoffman, Matthew D. et al. (2013). "Stochastic variational inference". In: *The Journal of Machine Learning Research* 14.1, pp. 1303–1347.

Holsti, Ole R. (2004). *Public opinion and American foreign policy*. University of Michigan Press.

Hopf, Ted (2002). *Social construction of international politics: identities & foreign policies, Moscow, 1955 and 1999*. Cornell University Press.

Horowitz, Michael C. and Allan C. Stam (2014). "How prior military experience influences the future militarized behavior of leaders". In: *International Organization* 68.03, pp. 527–559.

Humphrey, David C. (1984). "Tuesday lunch at the Johnson White House: A preliminary assessment". In: *Diplomatic History* 8.1, pp. 81–101.

Huysmans, Jef (2000). "The European Union and the securitization of migration". In: *JCMS: Journal of Common Market Studies* 38.5, pp. 751–777.

Ibrahim, Maggie (2005). "The securitization of migration: A racial discourse1". In: *International migration* 43.5, pp. 163–187.

Imai, Kosuke, James Lo, and Jonathan Olmsted (2016). "Fast estimation of ideal points with massive data". In: *American Political Science Review* 110.4, pp. 631–656.

Jaakkola, Tommi and Michael Jordan (1997). "A variational approach to Bayesian logistic regression models and their extensions". In: *Sixth International Workshop on Artificial Intelligence and Statistics*. Vol. 82, p. 4.

Jervis, Robert (1976). *Perception and misperception in international politics*. Princeton University Press.

— (1978). "Cooperation under the security dilemma". In: *World politics* 30.2, pp. 167–214.

Ji, Geng et al. (2017). "Bayesian Paragraph Vectors". In: *arXiv preprint arXiv:1711.03946*.

Johnson, Lyndon B. (1966). *State of the Union*.

Karnow, Stanley (1994). *Vietnam: A history*. Vol. 122. Random House.

King, Charles and Neil J. Melvin (2000). "Diaspora politics: ethnic linkages, foreign policy, and security in Eurasia". In: *International Security* 24.3, pp. 108–138.

Kocher, Matthew Adam, Thomas B. Pepinsky, and Stathis N. Kalyvas (2011). "Aerial bombing and counterinsurgency in the Vietnam War". In: *American Journal of Political Science* 55.2, pp. 201–218.

Kohl, Wilfred L. (2015). *French nuclear diplomacy*. Vol. 1422. Princeton University Press.

Kowert, Paul (2017). "Dimensions of Identity Construction and the Measurement of Differences among Countries". In: *Interpretive quantification: methodological explorations for critical and constructivist IR*. Ed. by J. Samuel Barkin and Laura Sjoberg. University of Michigan Press.

Kozlowski, Austin C., Matt Taddy, and James A. Evans (2018). "The Geometry of Culture: Analyzing Meaning through Word Embeddings". In: *arXiv preprint arXiv:1803.09288*.

Krebs, Ronald R. (2015). "How Dominant Narratives Rise and Fall: Military Conflict, Politics, and the Cold War Consensus". In: *International Organization* 69.04, pp. 809–845.

LaFeber, Walter (2013). *The New Cambridge History of American Foreign Relations: Volume 2, The American Search for Opportunity, 1865–1913*. Cambridge University Press.

Landauer, Thomas K. and Susan T. Dumais (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." In: *Psychological review* 104.2, p. 211.

Lauretig, Adam M. (2019). "Identification, Interpretability, and Bayesian Word Embeddings". In: *Proceedings of the Third Workshop on NLP and Computational Social Science*.

Lenz, Gabriel S. (2013). *Follow the leader?: how voters respond to politicians' policies and performance*. University of Chicago Press.

LeVeck, Brad L et al. (2014). "The role of self-interest in elite bargaining". In: *Proceedings of the National Academy of Sciences* 111.52, pp. 18536–18541.

Levy, Jack S (1997). "Prospect theory, rational choice, and international relations". In: *International Studies Quarterly* 41.1, pp. 87–112.

Levy, Omer and Yoav Goldberg (2014). "Dependency-based word embeddings". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 302–308.

Lowe, Will et al. (2011). "Scaling policy preferences from coded political texts". In: *Legislative studies quarterly* 36.1, pp. 123–155.

Lukes, Steven (2004). *Power: A radical view*. Macmillan International Higher Education.

MacKay, David J.C. and Radford M. Neal (1994). "Automatic relevance determination for neural networks". In: *Technical Report*. Cambridge University.

Malkasian, Carter (2004). "Toward a better understanding of attrition: The Korean and Vietnam Wars". In: *The Journal of Military History* 68.3, pp. 911–942.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of statistical natural language processing*. MIT press.

McAllister, William B. et al. (2015). *Toward "Thorough, Accurate, and Reliable": A History of the Foreign Relations of the United States Series*. Government Printing Office.

Mcauliffe, Jon D. and David M. Blei (2008). "Supervised topic models". In: *Advances in neural information processing systems*, pp. 121–128.

Mearsheimer, John J. (2001). *The tragedy of great power politics*. WW Norton & Company.

Michel, Marshall L. (1997). *Clashes: Air Combat over North Vietnam, 1965-1972*. Naval Institute Press.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations." In: *NAACL-HLT*. Vol. 13, pp. 746–751.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mimno, David and Andrew McCallum (2008). "Topic models conditioned on arbitrary features with dirichlet-multinomial regression". In: *UAI*, pp. 411–418.

Morgenthau, Hans Joachim (1973). *Politics among nations*. Knopf New York.

Mozer, Reagan et al. (2018). "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality". In: *arXiv preprint arXiv:1801.00644*.

Mueller, John and Mark G. Stewart (2012). "The terrorism delusion: America's overwrought response to September 11". In: *International Security* 37.1, pp. 81–110.

Mueller, John E. (1973). *War, presidents, and public opinion*. John Wiley & Sons.

Nelson, Laura K. (2017). "Computational Grounded Theory: A Methodological Framework". In: *Sociological Methods & Research*, p. 0049124117729703.

Nixon, Richard (1968). *Address Accepting the Presidential Nomination at the Republican National Convention in Miami Beach, Florida*.

— (1985). *No More Vietnams*. Simon and Schuster.

Norris, Clayton, Philip Schrodt, and John Beieler (2017). "PETRARCH2: Another event coding program". In: *The Journal of Open Source Software* 2.9.

Paterson, Thomas G. (1995). *Contesting Castro: The United States and the Triumph of the Cuban Revolution*. Oxford University Press, USA.

Pomeroy, Caleb, Niheer Dasandi, and Slava Jankin Mikhaylov (2018). "Multiplex Communities and the Emergence of International Conflict". In: *arXiv preprint arXiv:1806.00615*.

Porch, Douglas (2013). *Counterinsurgency: Exposing the myths of the new way of war*. Cambridge University Press.

Powell, Robert (2006). "War as a commitment problem". In: *International organization* 60.1, pp. 169–203.

Rabe, Stephen G. (1988). *Eisenhower and Latin America: The foreign policy of anticommunism*. UNC Press Books.

Rainey, Carlisle (2014). "Arguing for a negligible effect". In: *American Journal of Political Science* 58.4, pp. 1083–1091.

Ranganath, Rajesh et al. (2013). "An adaptive learning rate for stochastic variational inference". In: *International Conference on Machine Learning*, pp. 298–306.

Rheault, Ludovic et al. (2016). "Measuring emotion in parliamentary debates with automated textual analysis". In: *PloS one* 11.12, e0168843.

Ricks, Thomas E. (2012). "Let's Draft Our Kids". In: *The New York Times*.

Rivers, Douglas (2003). "Identification of multidimensional spatial voting models". In: *Typescript. Stanford University*.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (2016). "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association* 111.515, pp. 988–1003.

Robertson, Lt. Col. J. A. (2013). *Theater History of Operations Reports*. United States Air Force.

Rudolph, Maja et al. (2016). "Exponential family embeddings". In: *Advances in Neural Information Processing Systems*, pp. 478–486.

Rudolph, Maja et al. (2017). "Structured Embedding Models for Grouped Data". In: *Advances in Neural Information Processing Systems*, pp. 250–260.

Sagan, Scott D. (1997). "Why do states build nuclear weapons? Three models in search of a bomb". In: *International security* 21.3, pp. 54–86.

Salter, Mark B. and Can E. Mutlu (2013). *Research methods in critical security studies: An introduction*. Routledge.

Sarkees, Meredieth Reid and Frank Wayman (2010). *Resort to war: 1816-2007. Correlates of War*. Washington DC: CQ Press.

Saunders, Elizabeth N. (2011). *Leaders at war: how presidents shape military interventions*. Cornell University Press.

— (2017). "No Substitute for Experience: Presidents, Advisers, and Information in Group Decision Making". In: *International Organization* 71.1, pp. 219–247.

Sheffer, Lior et al. (2018). "Nonrepresentative representatives: an experimental study of the decision making of elected politicians". In: *American Political Science Review* 112.2, pp. 302–321.

Singer, J. David (1972). "The "Correlates of War" project: Interim report and rationale". In: *World Politics* 24.2, pp. 243–270.

Sorley, Lewis (1999). *A better war: The unexamined victories and final tragedy of America's last years in Vietnam*. Houghton Mifflin Harcourt.

State Department (2017). *About the Foreign Relations of the United States Series*.

Trachtenberg, Marc (2012). "The de Gaulle Problem". In: *Journal of Cold War Studies* 14.1, pp. 81–92.

Turney, Peter D. and Patrick Pantel (2010). "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* 37, pp. 141–188.

Valentino, Benjamin A., Paul K. Huth, and Sarah E. Croco (2010). "Bear any burden? How democracies minimize the costs of war". In: *The Journal of Politics* 72.2, pp. 528–544.

VanDeMark, Brian (1995). *Into the quagmire: Lyndon Johnson and the escalation of the Vietnam War*. Oxford University Press.

Wæver, Ole (1995). "Securitization and desecuritization". In: *On Security*. Columbia University Press, pp. 46–86.

Wainwright, Martin J. and Michael I. Jordan (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.

Wallach, Hanna M. (2006). "Topic modeling: beyond bag-of-words". In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 977–984.

Wallach, Hanna M., David M. Mimno, and Andrew McCallum (2009). "Rethinking LDA: Why priors matter". In: *Advances in neural information processing systems*, pp. 1973–1981.

Weeks, Jessica L. P. (2008). "Autocratic audience costs: Regime type and signaling resolve". In: *International Organization* 62.1, pp. 35–64.

— (2012). "Strongmen and straw men: Authoritarian regimes and the initiation of international conflict". In: *American Political Science Review* 106.02, pp. 326–347.

— (2014). *Dictators at war and peace*. Cornell University Press.

Weigley, Russell Frank (1977). *The American way of war: a history of United States military strategy and policy*. Indiana University Press.

Weisiger, Alex (2016). "Learning from the Battlefield: Information, Domestic Politics, and Interstate War Duration". In: *International Organization* 70.2, pp. 347–375.

Wendt, Alexander (1999). *Social theory of international politics*. Cambridge University Press.

Williams, Michael C. (2003). "Words, images, enemies: Securitization and international politics". In: *International studies quarterly* 47.4, pp. 511–531.

Wood, Simon N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Xu, Jiacheng and Greg Durrett (2018). "Spherical Latent Spaces for Stable Variational Autoencoders". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4503–4513.

Yarhi-Milo, Keren (2014). *Knowing the adversary: leaders, intelligence, and assessment of intentions in international relations*. Princeton University Press.

Zaller, John R. (1992). *The nature and origins of mass opinion*. Cambridge university press.

Zimmerman, William and Glenn Palmer (1983). "Words and deeds in Soviet foreign policy: The case of Soviet military expenditures". In: *American Political Science Review* 77.2, pp. 358–367.

Appendix A: Appendix: Elite Responses in Word and Deed to Casualties During the Vietnam War

A.1 Descriptive Statistics for Strategic and Tactical Bombing and Casualties

Casualties	Tactical Bombing Missions	Strategic Bombing Missions	Unemployment Rate	Inflation Rate	Nixon Dummy
Min. : 0.00	Min. : 0.0	Min. : 0	Min. :3.400	Min. :0.900	Min. :0.0000
1st Qu.: 2.00	1st Qu.: 71.0	1st Qu.:11	1st Qu.:3.800	1st Qu.:1.900	1st Qu.:0.0000
Median : 11.00	Median :165.0	Median :26	Median :4.400	Median :3.300	Median :0.0000
Mean : 17.47	Mean :173.4	Mean :27	Mean :4.546	Mean :3.692	Mean :0.4431
3rd Qu.: 25.00	3rd Qu.:257.0	3rd Qu.:42	3rd Qu.:5.400	3rd Qu.:5.100	3rd Qu.:1.0000
Max. :246.00	Max. :753.0	Max. :85	Max. :6.100	Max. :6.600	Max. :1.0000

A.2 Bombing at three timescales

Here, I display the number of bombing missions aggregated at the three time scales used in the paper (daily, weekly, and monthly).

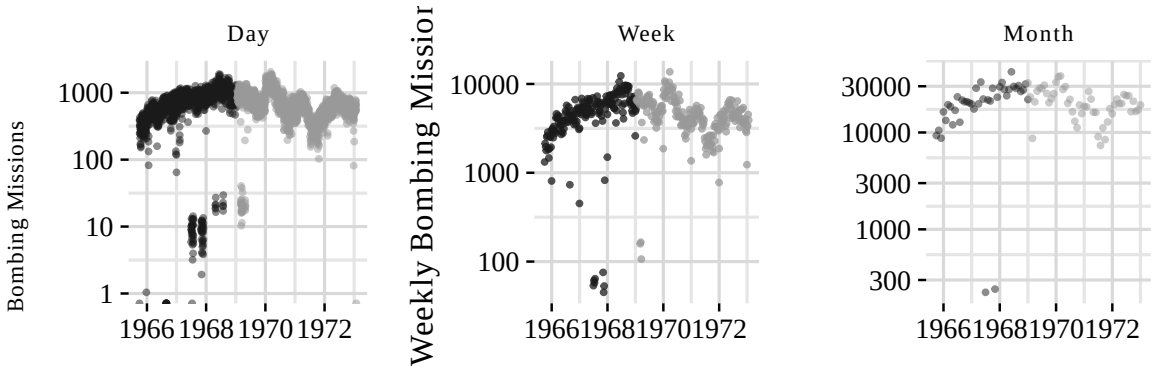

Figure A.1: Bombing Missions, 1965-1973. Dark gray is for missions during the Johnson administration, light gray is for missions during the Nixon administration.

A.3 Casualties at three timescales

Here, I show casualties at three timescales:

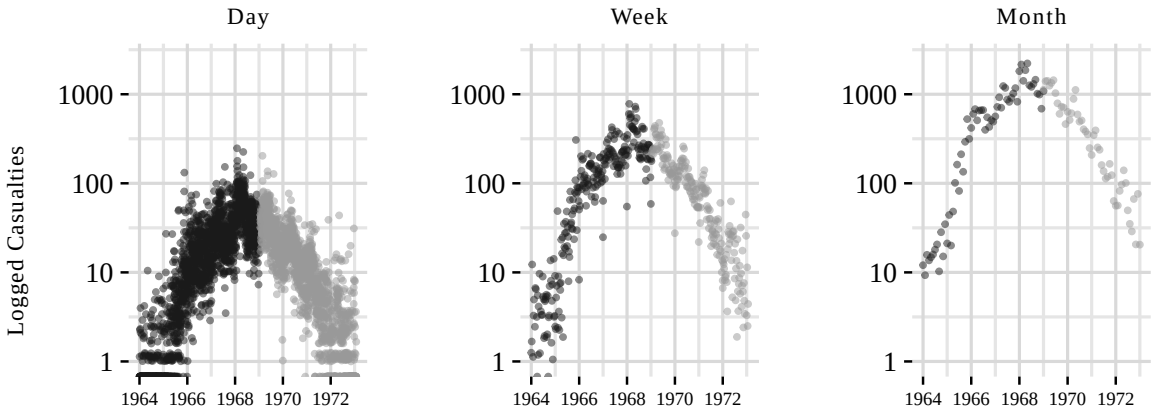

Figure A.2: Daily Casualties, 1964-1973. Dark gray is for casualties during the Johnson administration, light gray is for casualties during the Nixon administration.

A.4 The *Foreign Relations of the United States* Vietnam War Sub-corpus

Volumes in the *Foreign Relations of the United States* series are arranged according to country or theme by year. Here, I focus on documents from 1964 to 1973 in the Vietnam War FRUS collections, from the beginning of the year that marked the start of major conflict operations in Vietnam to the signing of the Paris Peace Accords on January 27, 1973, ending direct American involvement in Vietnam. These documents are particularly useful for scholars interested in elite decision-making: they represent an unfiltered view into private discussions among elites in "real time." Documents were extracted from the digital FRUS volumes, and tokenized, with stray letters, punctuation, and words associated with the archival nature of the corpus removed.

There are 61 types of documents in the FRUS Vietnam corpus, of which 25 are singletons. These types are scraped from the first two words of the document metadata, since this metadata appears in a consistent format, such as `Message from the Ambassador in Vietnam ( Lodge ) to the President`. In Figure A.3, the ten most frequently occurring document types are shown, and it is clear that internal, intra-elite communication comprises the majority of the documents in the corpus, and, based on their types, they represent private communication. Approximately 60% of the documents are memoranda, telegrams, or messages, with senders and receivers, addressing the creation and implementation of policy. A smaller number of the documents are summaries, reports, diary entries, or action memoranda, without a clear direction.

The "National Security" category refers to National Security Action Memoranda, National Security Studies, National Security Decisions, documents produced by the

National Security Council. The "Action Memorandum" category is also produced by members of the national security council, and con sender-receiver information in the metadata. The "Memorandum Prepared" category are mostly prepared by the Central Intelligence Agency, and do not contain directional information.

One category, "backchannel messages," stands out when compared to the seeming normality of other document types. These documents were communications between the White House (Nixon and Kissinger) and U.S. ambassadors, but kept secret from the official State Department bureaucracy. Similar backchannel messages were used during the Paris peace negotiations as well, and provide a more accurate view of policy decisions than "official" cables (*Foreign Relations of the United States Guide to Sources on Vietnam, 1969-1975*). The inclusion of these messages helps to assuage concerns that classification would remove all of the truly private communications between elites.

A.4.1 Describing the Documents

What do these documents look like, more generally? As Table A.1 shows, the corpus contains 4249 documents. The average document is 786 words long, and the standard deviation on this length is 886 words.

There is some heterogeneity in the distribution of document lengths: Johnson-era documents tend to be shorter and more numerous, while Nixon-era documents are fewer, but longer. This pattern is visible in Figures A.4 and A.5. In particular, 1972 has a noticeable increase in the length of documents, with a large number of longer outliers.

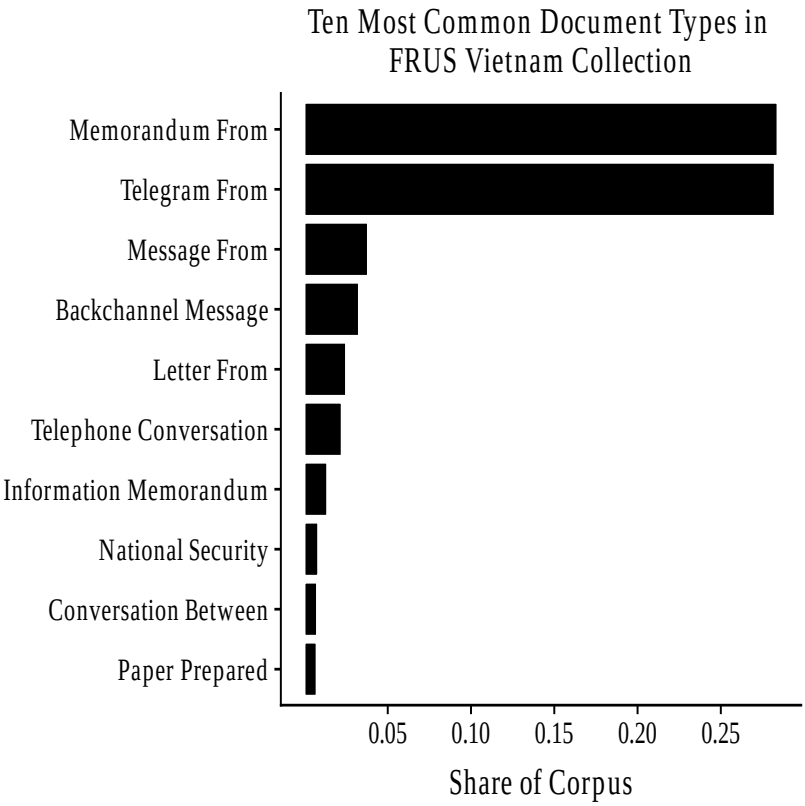

Figure A.3: 20 Most Frequent Document Types in the Foreign Relations of the United States Corpus

A.4.2 Network Structure of FRUS Documents

Communication Structure

The *FRUS* corpus contains information on the senders and receivers of a majority of the documents ($\approx 70\%$), which can help us understand decision-making processes during the war: who key actors are, and their relative importance. This helps us understand whose decisions we are analyzing. For example, while looking at the most common senders and receivers of documents in the corpus (in Figures A.6 and A.7) shows what we might expect: the President, National Security Adviser, and State Department are all high-frequency senders and receivers. Two names Figure A.6

Number of Documents	Average Document Length	Standard Deviation of Length	Shortest Document	Longest Document
4249 documents	775 words	876 words	78 words	16356 words

Table A.1: Corpus Summary Statistics

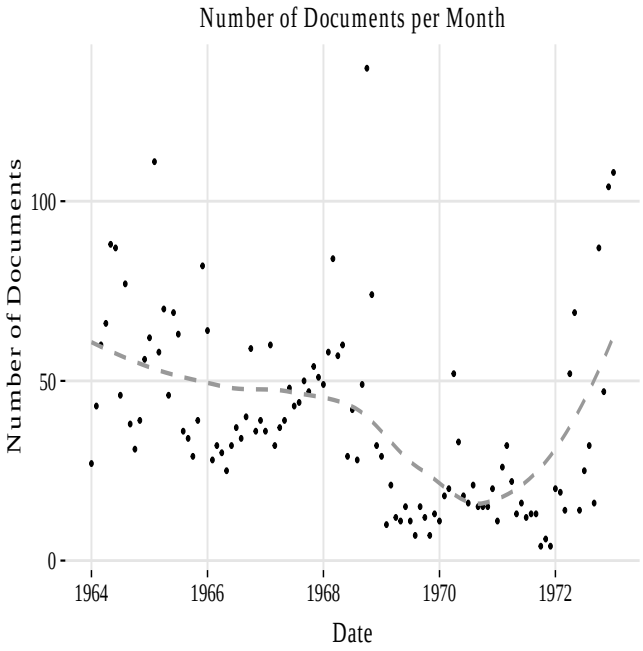

Figure A.4: Documents per Month

who are less well known are Michael Forrestal, a key advisor to McGeorge Bundy, who pushed for the 1963 coup in South Vietnam, and Senator Mike Mansfield, the Democratic Majority Leader, who opposed escalation during the Vietnam War.

However, when we instead look at the top ten senders and receivers in each year (Figures A.9 and A.10), a different pattern of behavior emerges. During the Johnson administration (1964-1968), the top senders were relatively evenly distributed among various members of the foreign policy apparatus: the State Department, President Johnson, and National Security Advisor Walt Rostow, among others. The French and Laotian embassies are also in the top ten senders and receivers during the Johnson

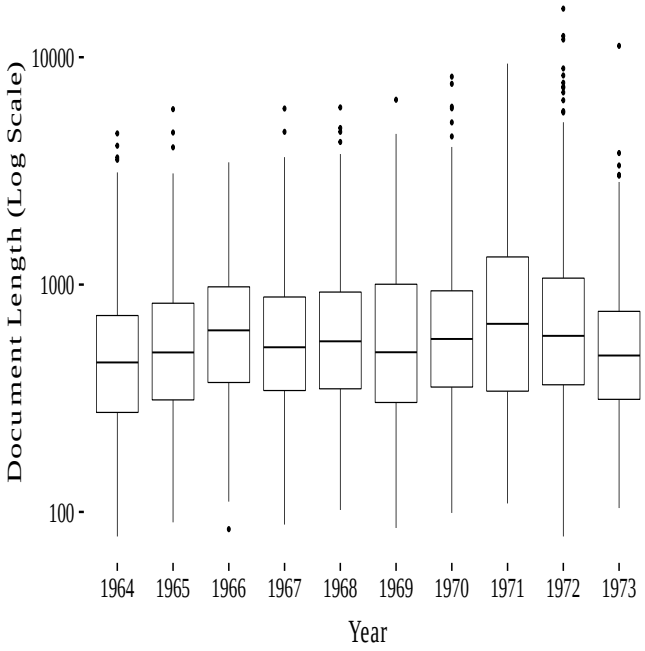

Figure A.5: Document Lengths Over Time

administration, which is unsurprising: Vietnam was a former French colony, and Laos neighbors Vietnam. With the start of the Nixon administration, communication is dominated by correspondence between President Nixon and National Security Advisor Henry Kissinger. Historians write that Nixon and Kissinger essentially ran a two-man foreign policy operation (Herring, 2008, ch. 17), and the extent to which they concentrated the foreign policy process during the Vietnam War stands out, particularly when compared to the more diffuse Johnson administration. This concentration of foreign policy with Nixon and Kissinger stands out when examining the top sender-receiver pairs in Figure A.8: "Kissinger to Nixon" is the second most frequently occurring dyad in the entire dataset. One concern, however, might be that historians are focusing on Kissinger because they retrospectively understand his significance to the White House: he appears central in these documents because historians

now believe he was central. This would be one case where having the entire universe of cables would be helpful: we would know if Kissinger *actually* communicated more, or if his cables are just over-sampled. Overall, this relational metadata allows us to examine the structure of foreign-policy decision-making, to understand who "key players" are in this corpus, and how this structure can contribute to foreign-policy outcomes (Saunders, 2017).

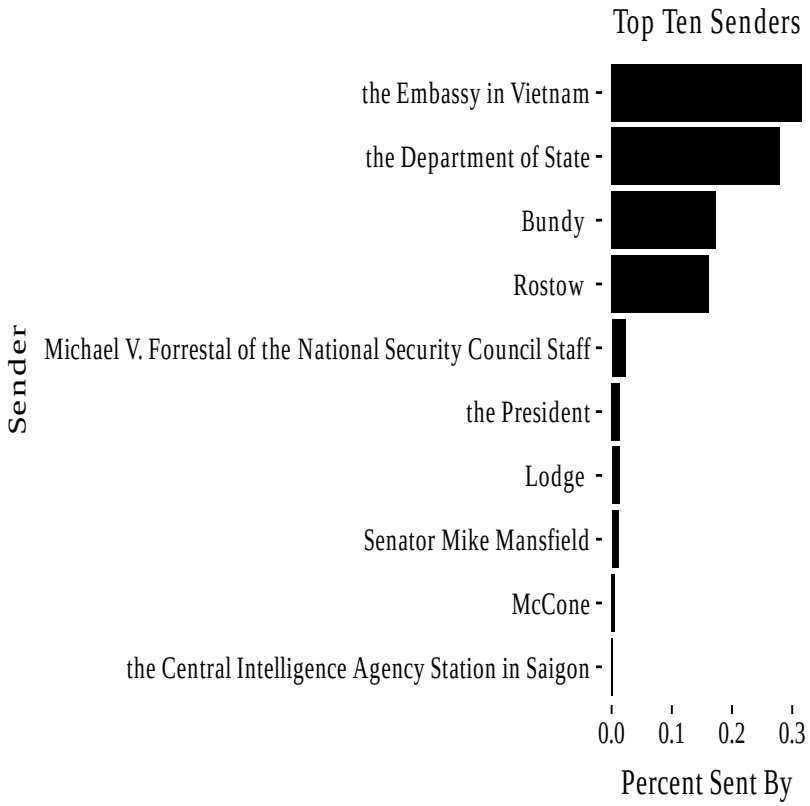

Figure A.6: Top Ten Senders

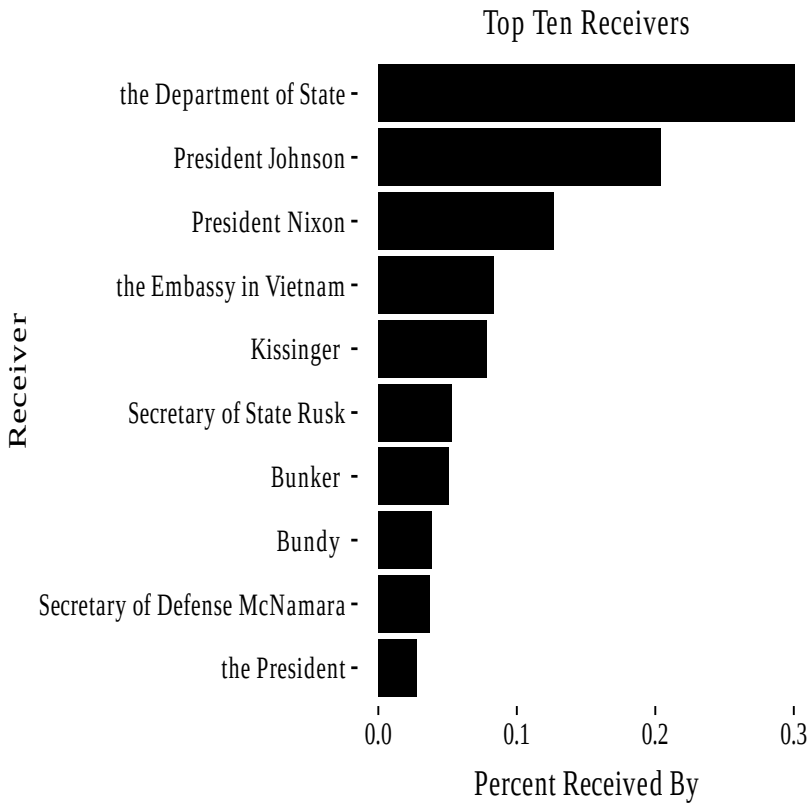

Figure A.7: Top Ten Receivers

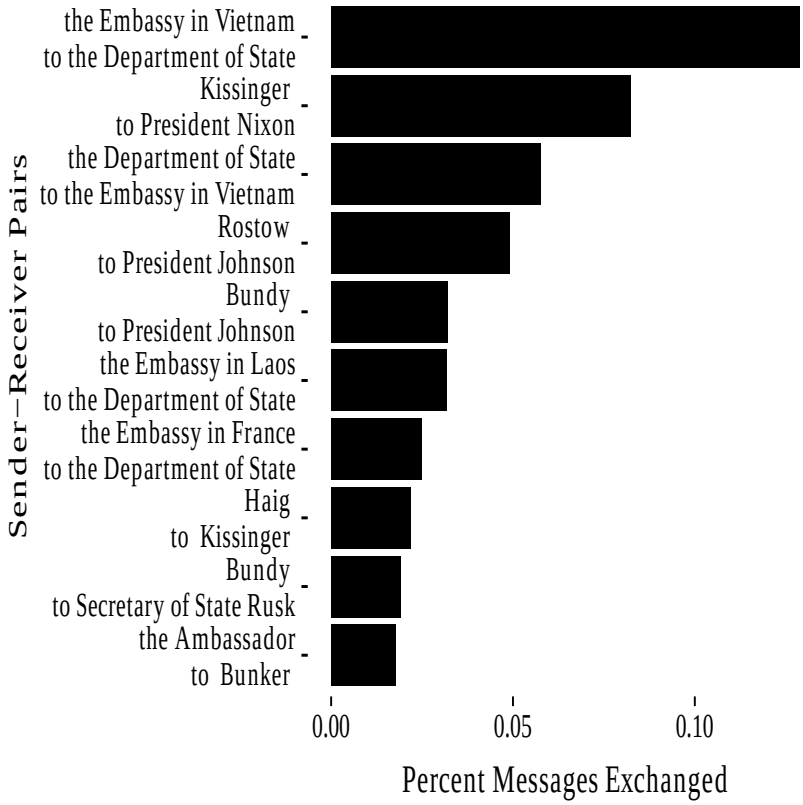

Figure A.8: Top Ten Sender-Receiver Pairs

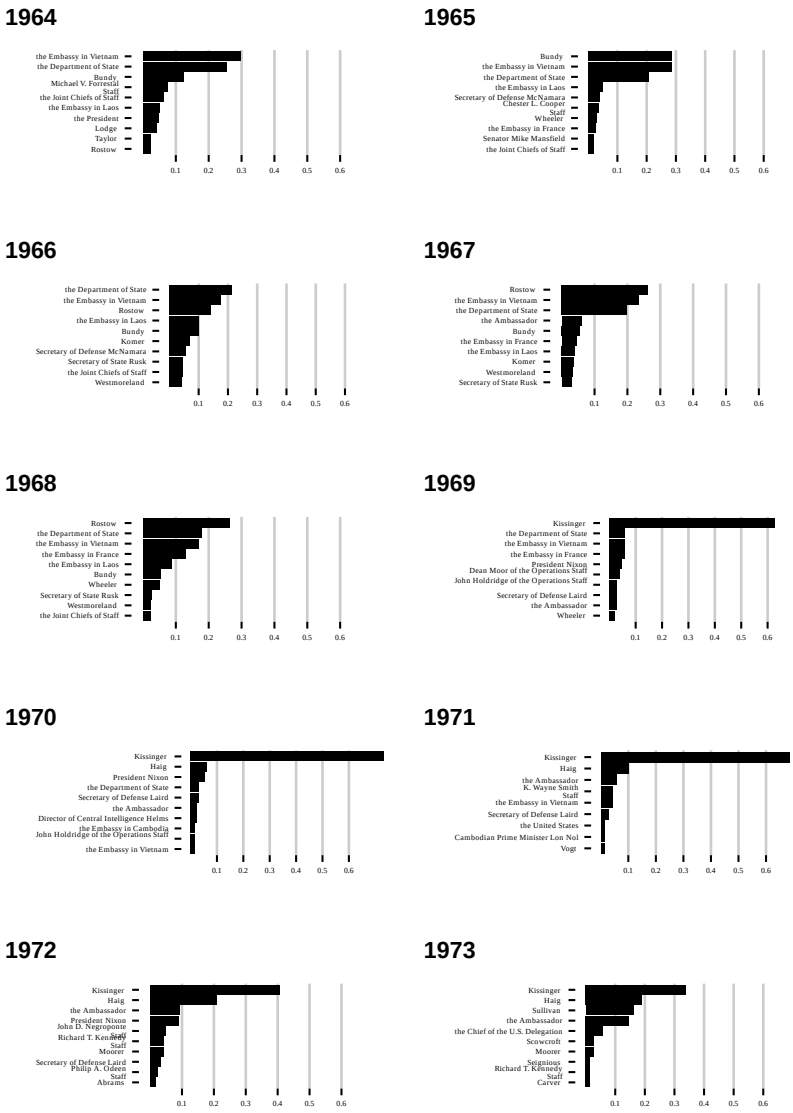

Figure A.9: Top Ten Senders by Year

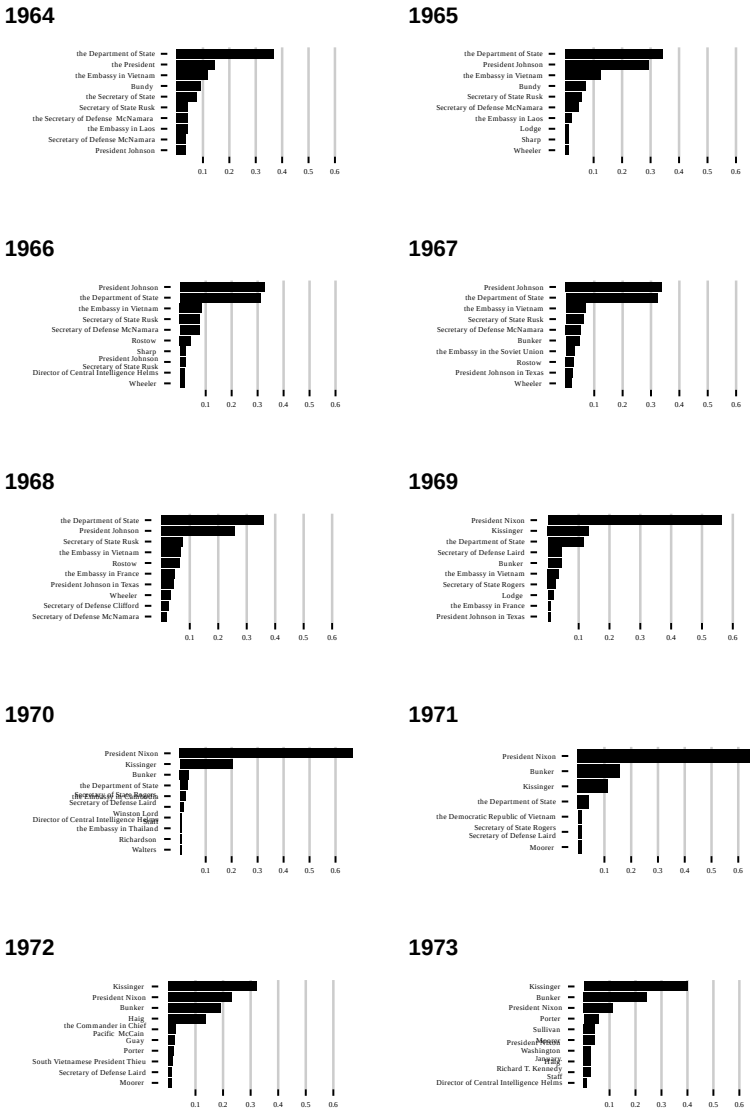

Figure A.10: Top Ten receivers by Year

A.5 Seven Day Lag of Casualties in the Political Model

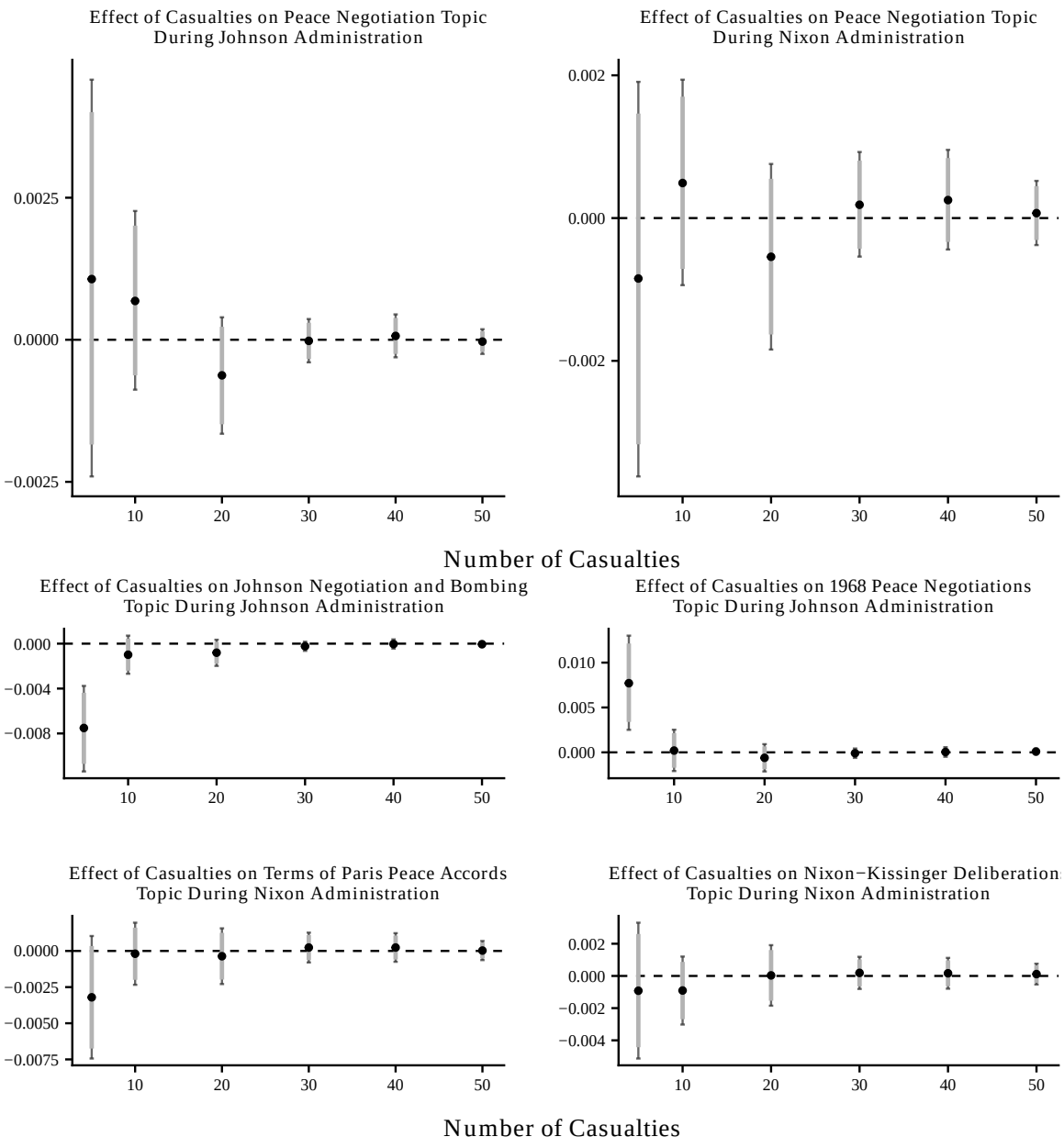

Figure A.11: Effect of an increase in casualties 7 days prior on the change in several negotiation topics across the Johnson and Nixon administrations. Plot shows both 90 and 95 percent confidence intervals.

A.6 One Day Lag of Casualties in the full model

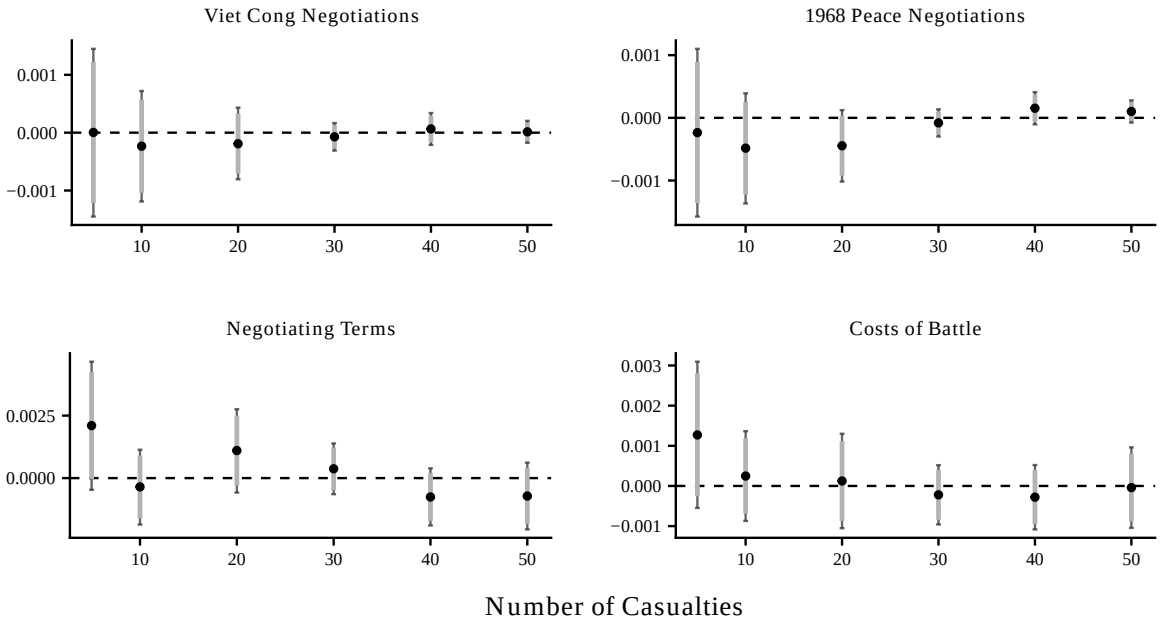

Figure A.12: Effect of an increase in casualties 7 days prior on the change in several negotiation topics across the Johnson and Nixon administrations. Plot shows both 90 and 95 percent confidence intervals.

Appendix B: Appendix: Identification, Interpretability, and Bayesian Word Embeddings

B.1 Comparing BWE to STM

Unlike the results from the Bayesian Word Embedding, the structural topic model detects no difference in topics before and after 1945. The top words, as determined by FREX score, are visible in B.1.

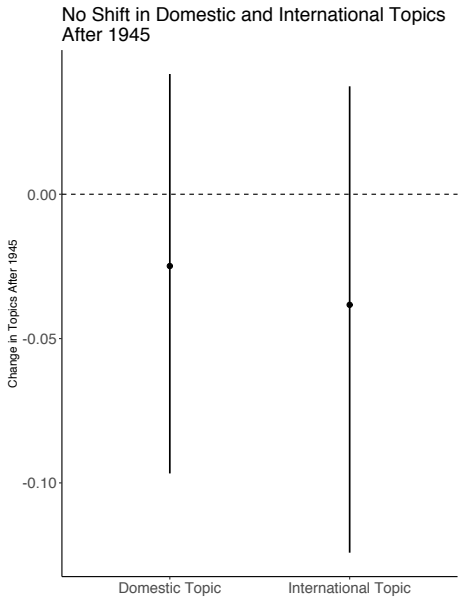

Figure B.1: There is no significant difference between foreign and international topics before and after 1945, uncertainty is displayed with 95% confidence intervals.

International Topic	Domestic Topic
representative	pacific
civilization	territory
making	question
international	whilst
tax	importance
popular	constitution
concern	slavery
supreme	domestic

Table B.1: Top eight words from structural topic model for international and domestic topics, by FREX score.

B.2 Inauguration Robustness Check

One concern with the role of internationalism in inaugural addresses is that by splitting at 1945, the "internationalism" of the pre-1945 sample is due to World War Two, and the Roosevelt presidency. To account for this, I re-estimate the Kolmogorov-Smirnov test from above, excluding the Roosevelt inaugural addresses, and present the results in B.2. This compares inaugurals from before 1932 to those after 1945, and the substantive results do not change.

B.3 GLM without Outliers

To ensure that the results presented in the main analysis were not the results of outliers, I removed any outliers and high-leverage points, and re-fit the model. The results were the same, as visible in Figure B.3.

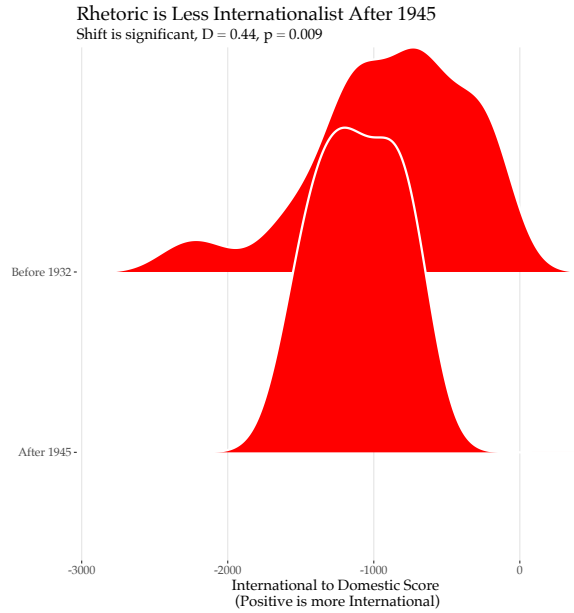

Figure B.2: Even excluding the Roosevelt administration, when only examining inaugural addresses from before 1932 and after 1945, the pre-1932 inaugural addresses are more internationalist than the post-1945 addresses.

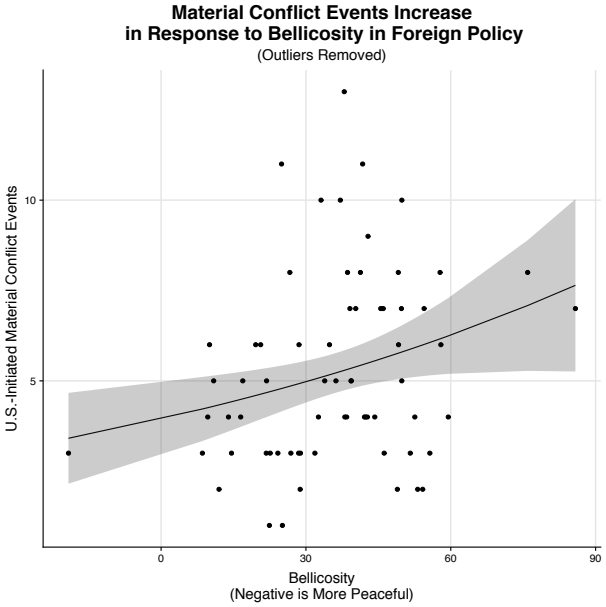

Figure B.3: An increase in bellicosity is associated with an increase in U.S.-initiated hostile events. The regression is from a Poisson generalized linear model, and uncertainty is displayed with 95% confidence intervals.

Appendix C: Appendix: Bayesian Word Embeddings, Securitization, and U.S. Foreign Policy

C.1 Alternate Link Functions

While I fit negative binomial models in the paper, there are other possibilities, especially given the large number of zeroes in the dependent variables. I pursue two strategies here, first, I fit a zero-inflated poisson Generalized Additive Model, which both models the probability of a non-zero observation, and the observed counts. I then dichotomize the outcome, and fit a logistic GAM.

C.2 Zero-inflated Poisson Regression

Here, I show the results for material conflict and cooperation across the three cases: Cuba, France, and China. Outcomes are predicted event counts. Note that, in the case of China, the material conflict model did not converge.

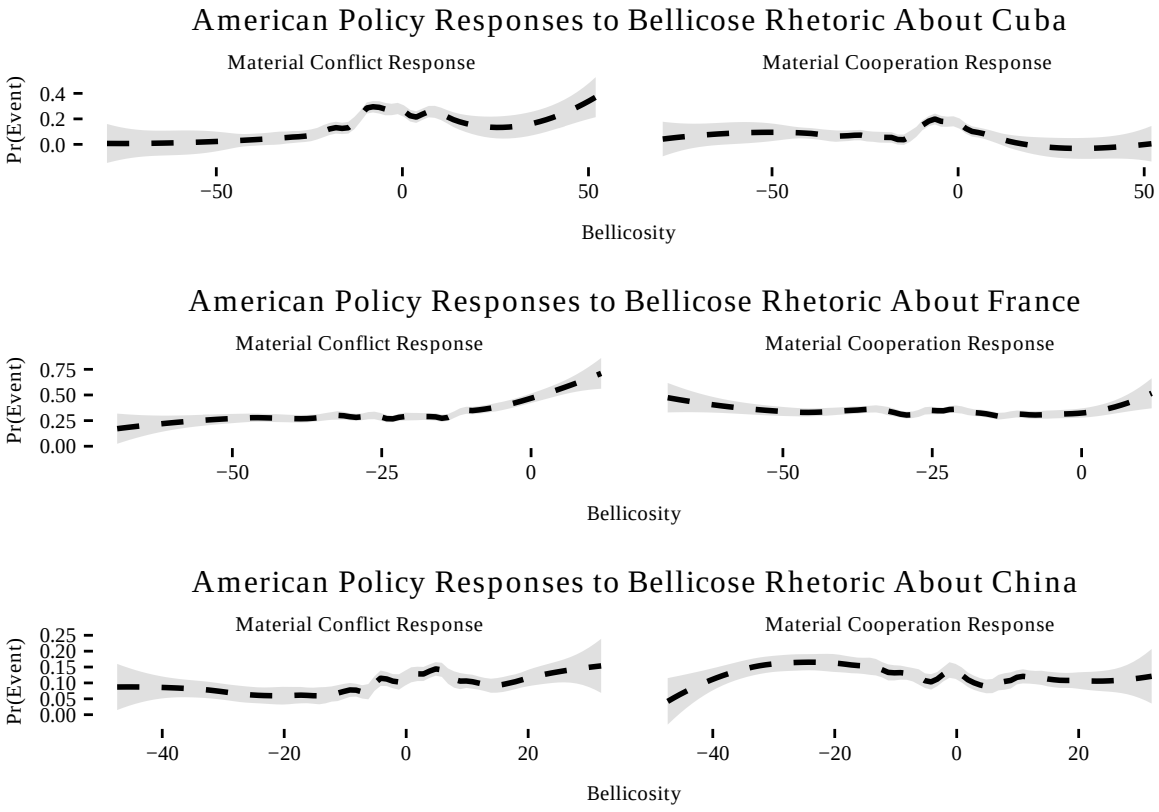

Figure C.1: Zero-inflated models, material conflict and cooperation.

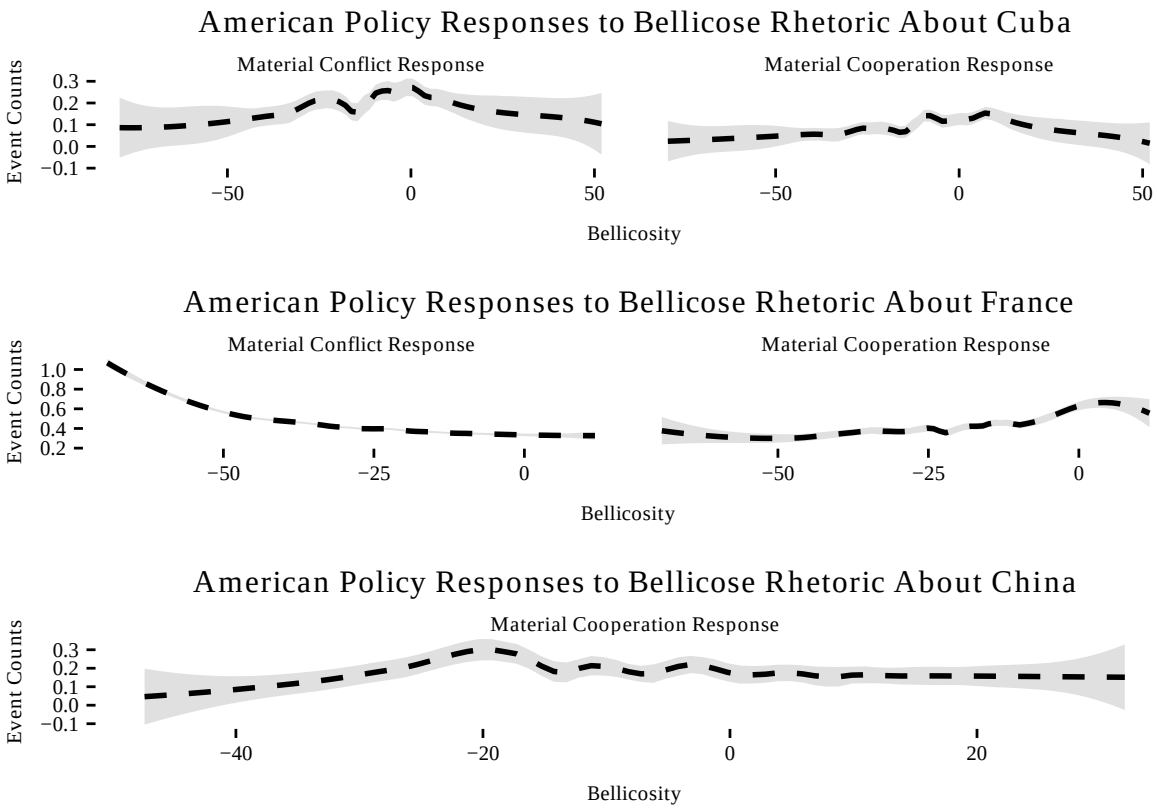

Figure C.2: Logistic regression models, material conflict and cooperation.

C.3 Logistic Regression

Here, I show the results from a logistic regression, after binarizing the dependent variable. Outcomes are predicted probabilities.