Automated Coding of Protest Event Data: Development and Applications

by

Alex Hanna

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Sociology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 6/7/2016

The dissertation is approved by the following members of the Final Oral Committee:
Pamela E. Oliver, Professor, Sociology
Chaeyoon Lim, Associate Professor, Sociology
Myra Marx Ferree, Professor, Sociology
Dhavan V. Shah, Professor, Journalism and Mass Communication
Xiaojin Zhu, Professor, Computer Science

ABSTRACT

Large-scale research of social movements has required more detailed, recent, and specific data about protest events. Analyses of these data allow for new insights into movement emergence, consequences, and tactical innovation and adaptation. One of the issues with this kind of analysis, however, is that the generation of event data is incredibly costly. Human coders must pore through news sources, looking for instances of protest and coding many variables by hand. Because of the high labor costs, projects are typically limited to one or two newspapers per country. This, in turn, exacerbates issues of selection and description biases.

This dissertation aims to address this issue with the development, validation, and application of a system for automating the generation of protest event data. This system, called the Machine-Learning Protest Event Data System (MPEDS), is the first of its kind coming from within the social movement community. MPEDS uses recent innovations from machine learning and natural language processing to generate protest event data with little to no human intervention. The system aims to have the effect of increasing the speed and reducing the labor costs associated with identifying and coding collective action events in news sources, thus increasing the timeliness of protest data and reducing biases due to excessive reliance on too few news sources. Work on MPEDS is ongoing, and to that end, the system will also be open, available for replication, and extendable by future social movement researchers, and social and computational scientists. By bringing cutting-edge computational tools to bear on a sociologically important set of questions, this dissertation has the potential to resolve longstanding data problems in the social movement field.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The idea to pursue this dissertation came during the GDELT Hackathon held at Penn State in 2013. I had been discussing the state-of-the-art in protest event data with Mohammed Idris when he suggested that I pursue work on an inchoate idea for using machine learning to generate new protest event data. About three years later, my work on this dissertation is finished, but the work on the idea is continuing to grow and flourish in a great community of scholars.

This dissertation could not have happened without the mentorship, love, and friendship of many, many people. The folks who have had the strongest hand in this work has been my dissertation committee, chaired by Pam Oliver. Pam has shaped me into a scholar and an academic since I began at Wisconsin. She is also one of the most kind-hearted and giving people I've ever met. I thank her for everything she's done for this dissertation and my work in general. This dissertation has also had an incredible amount of guidance from Chaeyoon Lim, Myra Marx Ferree, and Jerry Zhu, who kindly have discussed parts of it with me over the past several years. I also thank Dhavan V. Shah for being an excellent collaborator and someone who has forced me to think about how this work fits within larger media and communication processes. Other faculty at the University of Wisconsin who I'd like to thank include Mike Bell, Felix Elwert, Ivan Ermakoff, Lew Friedland, Young Mie Kim, Mike Massoglia, James Montgomery, Hernando Rojas, Robert Vargas, and Chris Wells.

At the University of Wisconsin, I've benefited immensely from my scholarly communities, including the Politics, Culture, and Society seminar, and the Sociology of Beer and Bourbon seminar. I also have benefited greatly from scholarly communities around the Bad Hessian blog, and sociology and computational social science colleagues on Twitter.

I want to thank all the research assistants I have worked with for the past several years. The graduate research assistants on this project, Katie Fallon and Emanuel Ubert, have proved invaluable. I've also had the pleasure with working with many great coders: Sam

Alhadeff, Omer Arain, Yijing Bai, Iakovos Balassi, Devon Betts, Anh Dang, Siying Fu, Stella Furlano, Taylor Johnson, Annie Johnston, Jesse Kearns, Sofi Lalonde, Allison Myren, Mat van Ommeren, Jodie Pope, Courtney Rodriguez, Erica Schultz, Malachy Schrobilgen, Charles Yeomans, Xiani Zhong, and Marilyn Zupkoff. The assistance of the wonderful folks at the Social Science Computing Cooperative – Dan Bongert, Ryan Horrisberger, and Nancy McDermott – and the Computer Sciences Lab – Tim Czerwonka – got the servers up and running and kept them running.

This research was supported by a National Science Foundation grant SES-1423784, an NSF Graduate Research Fellowship, and travel support from the University of Wisconsin-Madison Department of Sociology and Graduate School.

There are several scholars I want to thank for their engagement with this project and my research. Adam Slez, a former Wisconsin graduate student and now-faculty at University of Virginia, has served as a great mentor ever since we've met. Phil Schrodt is one of the pioneers in political event data, and he and his collaborators – John Beieler, Patrick Brandt, Tom Carsey, Andy Haltermann, Mohammed Idris, Erin Simpson, and Jay Yonamine – have been incredibly helpful in discussing all aspects of the event data generation process. Indeed, it was Phil who invited me to the hackathon which kicked off this dissertation in the first place.

Other folks I'd like to thank for dissertation and scholarly guidance include Richard Aviles, Chris Bail, Pablo Barberá, Neal Caren, Trey Causey, Rachel Einwohner, Deen Freelon, Peter Hart-Brinson, Andreas Jungherr, Matt Kearney, Brian Keegan, Brayden King, Hanspeter Kriesi, Benjamin Lind, Jasmine Lorenzini, Rahul Mahajan, Peter Marakov, Karissa McKelvey, Liz Merkhofer, Matt Moehr, Laura Nelson, Laura Norén, Brendan O'Connor, Sarah Soule, Junming Sui, Joshua Tucker, Ara Vartanian, Stefaan Walgrave, Dan Wang, Bruno Wueest, and JungHwan Yang.

I want to thank a great support network, inside the academe and out. Carlos Pereira has been my comrade-in-arms since our early reading group and activism days at Purdue. I also want to thank Taylan Acar, Lisa Brush, Brandon Gorman, David Calnitsky, Jenny Carlson, Clayton Childress, Barry Eidlin, Thomas Elliot, Tina Fetner, Kevin Gibbons, Colin Gillis, Adryan Glasgow, Eric Anthony Grollman, Amanda Hayden, Jill Hopke, Walker Kahn, Magda Konieczna, Jon Latner, Katie Lindstrom, Gina Longo, Aliza Luft, Neda Maghbouleh, Gina Neff, Molly Noble, Raul Pacheco-Vega, Adrienne Pagac, Madeleine Pape, Charles Peterson, Jen Schradie, Charles Seguin, Emma Shakeshaft, Gina Spitz, Zeynep Tufekci, Jaclyn Wypler, Katie Zaman, and Jienian Zhang.

I also have my families to celebrate. I thank my roller derby and my queer families. Those include the Mad Wreckin' Dolls and the (Season 12 champions!) the Quad Squad. Shoutouts to Block Ness Monster, Book It and Kit, Brokeback Jac, Chaotic Neutral, Conan the Librarian, Critical Tits, Dame Judi Bench, Dark Horse/Neigh Neigh, Girl Friday, Hammer Abby, Helena Skirt, Ironcide, Josie Simonis, Kel*Fire*Warrior, Kelly Quintavelle, La Bibliowrecka, Little Lebowski, Megalomaniac, Neon Arzach, Shantastic McAwesome, The Smacktivist, Siouxper Nova, Tear O'Bite, and Uno Mas. I also want to thank Gabe Javier, Katherine Charek Briggs, Liam Frahm-Gilles, Megan Milks, and Sara June Woods. Thanks to Anna Lauren Hoffman, my academic trans sister and a big data skeptic who pushes me to ruthlessly interrogate my methods of inquiry; Anders Zanichkowsky and Z! Haukeness, who I've worked with in the Dane County Trans Health Group; and Nicasio Andreas Reed, who has been one of my absolute best friends for the past three years. Nico reminds me every day that we're as varied as sea slugs – the one which looks like mold on cheese and the one which looks like Alexander McQueen Fall 2012 are both still sea slugs. And he reminds me to make kindness and awe the engines of my life.

Lastly, I thank my (biological) family: my mother, Angel, my sisters, Mariana and Sandy, my niece, Abriana, and nephews, Meshach, Bishoy, and Theo. My family has always emphasized the critical importance of education, and they have supported me every step of the way. Lastly, I thank Elvis the Wonder Cat for being a best bud for the past 10+ years.

INTRODUCTION

Protests and contentious collective action events more generally are an important part of democracy. Non-institutional outlets for the expression of collective grievances foster the exchange of ideas, build civil society organizations, and forge shared identities and understandings. When turned violent, however, protest can also become a major threat to democracy. Organized armed resistance threatens institutions and the rule of law within established territories.

The study of protests as rational behavior, undertaken by actors with concrete demands, motivations, and complex organizations – as opposed to an expression of irrational frustration or a maddening crowd – has only been around since the mid-60s, a consequence of myriad social movements of the time. Those first movement scholars focused primarily on protest *emergence* – what were the structural causes of protest activity, what grievances did those actors lodge, and against what were they struggling? Subsequent literatures focused on the professionalization of social movements and the arrangement of movements into fields. A next wave of work attempted to understand the cultural and discursive components of mobilization, asking what types of messages and "frames" were critical to motivate would-be activists into actual protesters. More recently, the social movement literature has taken seriously investigating the *outcomes* of movements. What are the consequences of movement activity? What kinds of protests – in terms of demographics, amount of resources, the structural position of activists in larger networks, and tactics and innovation in those tactics – tend to achieve their goals?

The larger scale interrogation of both protest emergence and consequences has necessitated the development of measures of protest events themselves. Moving away from the tradition of deep case study description, the initial movement scholars developed methodologies to define, identify, code, and quantify discrete instances of civil strife. Time-series analysis of frequencies of strikes, gatherings, rallies, and riots supplemented more traditional case studies. Cross-national and a new kind of long-term, over-time analysis became possible with quantitative information coded from multiple sources in several languages. In short, the social movement which produced the field of movement studies also advocated for and enacted a new mode of analyzing protest activity. This kind of analysis – which has come to be called "protest event analysis" – allows for new insights into movement emergence and outcomes. The analyses and the data which accompanies them allows for variation across time and space and for innovations in quantitative methodologies to be applied to the study of movements.

One of the issues with this kind of analysis, however, is that the generation of event data is incredibly costly. Human coders must pore through historical sources, looking for instances of protest within these sources and coding many variables by hand. For this reason, projects are typically limited to one or two newspapers per country. And even then, this is an enormous task. Another issue is that because we are limited to one source, there is a larger chance that selection biases of a particular newspaper will play a larger role in shaping the dataset.

This dissertation grew from a concern with the lack of availability of protest event data for cross-national research and for more recent time periods. Two projects inspired this investigation. The first project revisited the relationship of grievances and protests. A collaborator (Chaeyoon Lim) and I wanted to investigate the cross-national relationship between grievances, public interpretations of grievances, and protest activity on a large-scale cross-national basis. We planned to use two protest event datasets: the Cross-National Time-Series (CNTS) dataset and the then-newly released Global Database of Events, Language, and Tone (GDELT). However, when we started to dig into these data, they proved to be lacking for our purposes. Serious questions arose about their data quality, the news sources on which they were based, and the opaqueness of their data generation processes.

A second project, my original dissertation project, was to be an exploration of discourses around democratic politics in Egypt from 2003 to 2010 in the run-up to the 2011 Egyptian revolution, and how social media discourses may have slowly become more mainstream within larger Egyptian broadcast media. A second portion of the project was to be an interrogation of the relationship between the rise of democratic discourses and collective action events, especially from the years just before the revolution. However, the lack of suitable protest event data (along with the lack of travel funding) doomed that project.

This began a quest to dig into the field of protest event analysis, and more specifically, automated event data creation. In conversations with other event data analysts it then dawned upon me that most automated projects paid insufficient attention to the generation of protest event data. These projects tended to be focused on interstate conflict or conflicts between the state and armed militias. Protest was one interaction form of many, but because of the different substantive focuses, these data were not sufficient to address questions relevant to many movement scholars.

This dissertation, then, focuses on the history, development, validation, and application of an automated protest event data system. This system, called the Machine-Learning Protest Event Data System, is the first of its kind coming from within the social movement community. The availability of accessible machine learning methods and tools to train new models made this project a reality in a way that would have been more difficult ten or even five years ago. This dissertation outlines the history of political event data writ large, traces the development of a new system for protest event data generation, and applies to it to a modern social movement.

In Chapter 1, I trace two lineages of political event data within political science and sociology. The two branches of event data began with similar origins, but diverged into two branches. The first branch became concerned with international conflict and political interactions. The data was typically dyadic in nature, between either two nation-states or one nation-state and some organized challenger. The other branch, however, delved into protests and protesters themselves, focusing on repertoires of contention, the wide variety of claims being made by protest actors, and the organizations involved with the event of interest. I compare two datasets from these respective lineages: the Social, Political and

Economic Event Database (SPEED) in the political interaction tradition and the Dynamics of Collective Action (DoCA) dataset in the social movement one. In a quantitative re-analysis of Doug McAdam's seminal book (1982) on the Civil Rights Movement, I highlight the different attention to time trends, repertoires of contention, protest claims, and protest size given in each dataset. This then sets the stage for the creation of automated event data systems tailored to social movements research.

In Chapter 2, I build, test, and validate the MPEDS system in earnest. The Machine-Learning Protest Event Data System is able to code protest events from any electronically-available, English-language news source using advances from natural language processing and machine learning. Such a system should have the effect of increasing the speed and reducing the labor costs associated with identifying and coding collective action in news sources, thus increasing the timeliness of protest data and reducing biases due to excessive reliance on too few news sources. I describe the different stages necessary for the development of the system, including the three tasks – the haystack task, the closed-ended coding task, and the open-ended coding task. I also compare different sources of training data and their suitability for use in training the classifiers. I then validate the system and find high accuracy with most of its components. Work on MPEDS is ongoing, and to that end, the system is also be open, available for replication, and extendable by future social movement researchers, and social and computational scientists.

In the final chapter, Chapter 3, I apply the MPEDS system to a modern social movement, the Black Lives Matter movement. I first theoretically outline how many Black Lives Matter protests can be construed as backlash mobilization in the face of intense Black political repression. Following Oliver (2008), I describe how protest policing cannot be divorced from regular policing, and that mass incarceration, policing, and surveillance of Black neighborhoods constitutes Black political repression. I then take a media ecology approach to describing the media attention given to the Black Lives Matter protests and develop a model of the media ecology in operation within backlash mobilization. Using national, international, and local broadcast news media, and Twitter data, I assess research questions about the operation of the media ecology in the days following a repression event, namely, police killings and police (non-)indictment announcements. I find that local news media is still important in the media ecology for both attention to a particular police killing and the coverage of the protests which follow it. I conclude with a discussion of the necessity to look at local news media in movement research, especially in digitally-enabled movements like Black Lives Matter.

Before getting into the meat of this dissertation, I would like to offer a word on positionality, that is, how who I am informs what I see (and don't see) in my analyses, how I relate to the subjects of my research, and who the research is actually for. It isn't terribly typical for quantitative social scientists to discuss their positionality with respect to their research subject area. But as a firm proponent of feminist social science (e.g. Harding, 1987) and with an eye towards the burgeoning literature on critical data studies, I feel that it is necessary to answer the questions of who I am and who this research is for. I am an Egyptian-American transgender person and woman of color. I am able-bodied and middle-class. I am queer and the children of immigrants. Those facts alone do not make me a very likely suspect to be involved in computational social science and data science analysis, which is typically dominated by cisgender, straight men. Because of that background, I do hope to offer a more critical and reflexive eye towards the data generation process in political event data analysis. But I know it also means that I may be missing elements in my analyses related to protests which are undertaken by certain opposed groups with whom I do not identify, such as Black, lower-class, and disabled groups. This lack of insight may also extend to movements of privileged groups, of those who are White or affluent. This research is also undertaken for the purpose of enriching movement scholarship at large. Movement scholars have diverse aims, from scholarly engagement to community outreach and advocacy. I hope that this research would serve the purposes of both ends of that spectrum.

Furthermore, I am attentive that the last chapter focuses on Black movements and violence against Black people. I have been marginally involved with anti-racist activities in Madison, Wisconsin, with both the mobilization around the death of Tony Robinson and supporting Black trans women who have been incarcerated in Dane County. But I am also not "in" the movement, and I am a non-Black person of color. Because of that, I am less familiar with the priorities of that movement than I suspect I should be. I also do not experience the same kind of oppressions that Black people do in the United States. That said, this work would benefit greatly from more intensive dialog with activists in this movement. I hope that this work presents itself to be useful and insightful to those who are involved.

1 THE TWO LINEAGES OF CONTENTIOUS POLITICAL EVENT DATA

Protest event data has a long tradition within social movement scholarship. The founding members of the subfield pioneered the systematic enumeration of contentious events decades ago. The movements on which they focused are canonical: the Civil Rights Movement, New Social Movements in Europe, and the rise of the American Labor. There is another tradition of event data, however, with which the protest event data tradition has not had much interaction. That event data is typically related to war, conflict, and political interactions between states. In the beginning of both of these fields, the delineation between protest event data and conflict data was much more amorphous and the boundaries were much less strict than they are today. As these fields became more and more specialized, however, the two lineages drew further away.

While the specializations were informed by research priorities of each discipline, recently we have seen a rising interest in each subfield looking towards data from the other. Some of the more recent political interaction projects have taken an explicit interest in protest events and have sought to integrate them into their systems. Meanwhile, movement scholars and political sociologists are taking an interest in investigating collective violence and genocide. It is then important to understand the underlying structures and assumptions of these two data lineages.

The purpose of this paper is to analyze the history and use of data from these two areas of study, which I call the protest event data and political interaction data branches. While protest event data has been generated primarily by social movement scholars and sociologists, political interaction data has been generated by political scientists and conflict scholars. I describe the lineages of these two branches of event data, from their inceptions in war studies to modern developments which use computer-aided technologies. In these histories, I aim to detail multiple facets of the data-generation process of these data, including the source texts, the retrieval process, their coder training and interfacing, software used for annotation, and the details of the final product. I then compare how protest event data analysis and political interaction data systems treat several variables of interest: repertoires of contention, movement claims, initiating groups and organizations, and protest targets. Lastly, I compare two datasets – the Dynamics of Collective Action (DoCA) dataset and the Social, Political, and Economic Event Database (SPEED) – against a seminal text in the social movements tradition, Doug McAdam's *Political Process and the Development of the Black Insurgency, 1930 - 1970* (1982). I conclude with a discussion of future directions for both lineages, and fruitful points of interaction.

Furthermore, this paper provides the motivation for the development of the Machine-Learning Protest Event Data System (MPEDS), an automated system which generates protest event data with minimal human intervention using computational methods, namely machine learning and natural language processing. By bringing computational tools to generate data that are at the crux of social movement research, MPEDS's methodological innovation has the potential to resolve longstanding data problems in this field.

1.1 Lineages of Political Event Data

> We are equally grateful to that army of historians who alternately delighted us with their painstakingly gathered military statistics, and infuriated us with their Olympian disdain for the comparative and the nomothetic. – Singer and Small (1972, Acknowledgements)

Political interaction / conflict data[1] and protest event data share a lineage which can be traced back to common roots in a small group of political scientists, sociologists, and historians. From these common roots emerged two parallel but very different branches of data collection and social scientific inquiry. The first branch comes primarily out of political science, namely the subfields of international relations and conflict studies. This

[1] In this paper, I will use the terms "political interaction data" and "conflict data" somewhat interchangeably. While some datasets focus on solely conflict, these two types of data are part of the same tradition, as noted below.

body of work focuses primarily on inter- and intrastate conflict, including but not limited to interstate war, civil war, and both non-violent and violent challenges to state power. These datasets tend to enumerate the details of wars, armed conflicts, and challenges to state power, including variables relating to frequency, intensity, duration, and involved parties. The second branch concerns contentious gatherings and activity, mainly emerging from sociology and history, with the occasional political scientist, and the subfields of social movements, comparative-historical sociology, and social science history. This work and its datasets center around protests in many different forms (including marches, rallies, press conferences, and petition signing), collective violence and rioting, and strikes. I show in this section that the conceptual division between these two branches has had a significant impact on the shape of their methodological developments, including their primary sources, their data gathering strategies, the accepted ontologies of actors and events, and what has qualified as data. More recently, these two branches have seen some productive synthesis, especially with the development of machine-coded and machine-assisted approaches, which may indicate future fruitful endeavors for cross-discipline collaboration.

The earliest quantitative chroniclers of both political interactions and protest events were those concerned with enumerating and characterizing the magnitude of war and conflict. Sorokin (1937) and his compendium of *Social and Cultural Dynamics* are often attributed with being one of the founding members of this group. Sorokin compiled his collections from news reports and local histories, and took a special interest in building statistical scales which quantified the magnitude and intensity of these conflicts. His initial efforts were joined by updated statistical handbooks of Wright (1942) and Richardson (1960). The World Data Analysis Program at Yale University initiated the *World Handbook of Political Social Indicators*, the first edition of which was created in the initial heyday of both branches (Russett et al., 1964), and has been continually updated with second (Taylor and Hudson, 1972) and third (Taylor and Jodice, 1983) editions.

Political interaction event data: Conflict, cooperation, and mediation

The first maturation and systemization of major interstate conflict data appeared with David Singer's publication of the Correlates of War (COW) project (Singer, 1972; Singer and Small, 1972). COW focuses on wars as episodic events, which means it frames conflict in terms of a broader war rather than focusing on individual battles and actions. A "war" is defined as a prolonged military engagement with at least 1,000 battle deaths between two officially recognized armies and at least 100 deaths on each side. (Yonamine, 2013, pp. 4). COW is sourced from records of newspapers, books, microfilms, and online sources. Each conflict is then coded for the states involved, the entry and exit of each state into the conflict, the duration of the conflict, the location of the conflict, and the number of battle deaths. Continuing in the tradition of Sorokin, COW also attempts to quantify the magnitude of the conflict by calculating nation-months, the proportion of battle deaths compared to the total number of armed forces per country, and the proportion of battle deaths compared to the population of the country. In its 1972 edition, COW quantified some 200 conflicts from 1861 to 1965.

A parallel but markedly different effort to quantify conflict had been undertaken by McClelland and the World Event/Interaction Survey (McClelland and Hoggard, 1968; McClelland, 1999). His focus was not the enumeration of wars but instead a broader, day-to-day understanding of interactions between nation-states, conflictual and otherwise. Events in the WEIS ontology span 63 different categories that can be characterized on a cooperation-conflict continuum: verbal cooperation, cooperative action, participation, verbal conflict – defensive, verbal conflict – offensive, and lastly conflict action. One of the major innovations of WEIS was a move from focusing on *episodic* conflict, to *discrete* and *atomic* actions undertaken by nation-states (Yonamine, 2013; Schrodt, 2012*b*). These events were specific, daily-level events such as attacks, denunciations of leaders, and extending economic aid (McClelland and Hoggard, 1968, pp. 714-15). In latest dataset available on ICPSR (McClelland, 1999), events are sourced from the New York Times Index (NYTI) and are coded for a sparse set of variables: actor, target, action, date, and arena (of the world). There are 91,240 coded events measured from 1966-1978, a huge increase from the nearly 200 conflicts noted in COW.

Growth within the development of conflict datasets within the next few years focused on expanding the definition of conflict and the types of events for consideration. The team which produced COW moved from an exclusive focus on war to an expanded focus on any kind of military interaction between states. This expanded focus birthed the Militarized Interstate Disputes (MID) dataset (Jones, Bremer and Singer, 1996). WEIS's event ontology was expanded by Edward Azar's Conflict and Peace Data Bank (COPDAB) (Azar, 1980). Both of these datasets continued using similar data sources for enumeration, MID being sourced from government documents, historical monographs, case studies, diplomatical histories, and newspapers, many of them in multiple languages. COPDAB expanded from the single source of the NYTI to include newspapers and handbooks of multiple origins. MID structured its episodes as a series of interactions between states, most of which end up in war. COPDAB expanded the number of variables by adding measures of scale (e.g. how conflictual or cooperative is the event).

In the early 1990s, a parallel program of conflict data emerged at Uppsala University in Sweden and the Peace Research Institute Oslo (PRIO), culminating in the Uppsala Conflict Data Program (UCDP). UCDP's original program focused on episodic conflicts from 1989 to 1992, characterizing the conflicts with a much lower threshold (25 battle deaths) than COW (Wallensteen and Axell, 1994). The 1994 UCDP dataset uses a variety of sources, typically major English newspapers within a global scope. Updates to the UCDP dataset have backcoded it to 1946, and the program releases data on new conflicts every year, now updated to 2014 (Pettersson and Wallensteen, 2015).

A more major revision to the UCDP/PRIO's episodic data came with the development of the Armed Conflict Location and Event Dataset (ACLED) (Raleigh et al., 2010). ACLED focuses on eight types of violence: four battle-focused categories, a non-violent conflict event, rioting/protesting, violence against civilians, and non-violent transfer of location control. ACLED is one of the first conflict datasets which moves beyond focusing on states as a necessary political actor and also allows rebel groups and other political challengers to feature in conflict actions. It specifies the location of the event on a substate level, allowing for a more geographically granular approach to studying conflict. The intent behind this geographical disaggregation is to study internal conflicts, such as civil wars and conflict between non-state actors (e.g. militias) which do not involve any state intervention or state forces.

The development of automated methods for generating political interaction and conflict data proved to be a major turn during the 1990s. The Kansas Event Data System (KEDS) was the first major incursion into the development of automated methods (Schrodt, Davis and Weddle, 1994; Gerner and Schrodt, 1994). KEDS (and its successors) uses shallow parsing of text, combined with actor and event dictionaries, to produce interaction data on discrete political actions. The combined actor and event dictionaries are called an *ontology*. KEDS identifies subject-verb-object triads within the text, matches the subjects and objects to the actor dictionary and the verbs to the event dictionary. The original KEDS system used the WEIS categories. It incorporates all the relevant variables from WEIS and COPDAB, that is, the actor-target dyad, the event category, and an index of cooperative-conflict. Instead of using the multiple primary sources of COPDAB or the single primary source of the NYTI, KEDS relies primarily on newswire services, namely Reuters and the Agence-France Press (AFP). Gerner and Schrodt (1994) found that using only Reuters, they were able to track the dynamics of events between several country dyads in the Middle East with similar results to WEIS.

KEDS and Phil Schrodt have set the stage for an impressive amount of follow-up work in the automation of political event data, including the development of new actor/event ontologies, political event datasets, and new software for the coding of news sources. Gerner et al. (2002) created an ontology – the Conflict and Mediation Event Observations (CAMEO)

– focused on studying third-party mediation in international disputes. Bond and Bond (1995) created the Protocol for the Assessment of Nonviolent Direct Action (PANDA), oriented to study nonviolent direct action, including protest and demonstrations. Its authors went further to develop the Integrated Data for Events Analysis (IDEA) ontology (Bond et al., 2003). Although these ontologies had an explicit focus on collective action, few movement scholars have used them. The number of datasets which grew from KEDS include Schrodt's own Levant datasets (Schrodt, 2015*a*,*b*), the 10 Million International Dyadic dataset (King and Lowe, 2008), the Integrated Crisis Event Warning System (ICEWS) data (O'Brien, 2010; Boschee et al., 2015), the Global Data on Events, Location, and Tone (GDELT) data (Leetaru and Schrodt, 2013), and a fourth edition in the *World Handbook* series (Jenkins et al., N.d.). The software which has been improved upon or adapted from KEDS include TABARI (an early rewrite of the original KEDS software), PETRARCH (a modern effort to convert TABARI to Python and incorporate a fuller event data pipeline), JABARI (a proprietary version of TABARI owned by Lockheed Martin and implemented in Java), and VRA Reader (another proprietary implementation based off of the IDEA ontology).

Apart from the KEDS branch of the political interaction event tree, additional datasets and systems are being created with and without the aid of automated methods. The Social, Political, and Economic Event Database (SPEED) project aims to provide a richer dataset on political violence, mass expression (including protest), and disruptive state acts. They use a semi-automated system which processes stories through a machine learning based filter, which are then presented to humans for manual annotation (Nardulli, Althaus and Hayes, 2015). The Mass Mobilization on Autocracies Database (MMAD) records data on mass mobilization events in dictatorships (Rød and Weidmann, 2014). They also use a machine learning screening process to filter out non-relevant events (Croicu and Weidmann, 2015).

These datasets serve the priorities of major subfields within political science, namely conflict processes, peace studies, international relations, and foreign policy. The data are primarily concerned with nation-state actors, or non-state actors (armed and otherwise) who pose a serious political threat to the established state order. Research questions using these data tend to focus on the probability of inter- and intrastate conflict, national orientations towards cooperation and remediation, and the stability of single states. These research agendas and disciplinary orientations, thus, are in sharp distinction with those of social movement and protest event studies.

Protest event data

The history of protest event data starts nearly at the same time that Singer and McClelland were creating COW and WEIS. Rucht, Koopmans and Neidhardt (1999) and Hutter (2014*a*) locate the initiators of protest event data within the *World Handbook* series. Tarrow's (1999) identifies four trends in protest event data, two of which I focus on: the *events-in-history* and *event histories* approaches. The events-in-history approach is a departure from "event-ful" histories of sociologists in the comparative-historical tradition, especially William H. Sewell Jr. (1980; 1996). This approach began with the group of scholars coming out of Center for International Studies at Princeton directed by Harry Eckstein, including Charles Tilly and Ted Gurr (Franzosi, 2004, pp. 37-39). The focus of the Center had been civil strife and collective violence, which culminated in Eckstein's own work on a range of conflict activities within single polities (1964), Gurr's landmark work on the relative deprivation explanation of political conflict (1970), and the series of studies by Tilly on collective violence and collective action (Tilly and Rule, 1965; Tilly, Tilly and Tilly, 1975; Tilly, 1978).

The events-in-history approach focuses on "historical sequences of events in relation to one another in particular historical configurations" (Tarrow, 1999, pp. 48) and "the relationship of events to other events, the organizations and actors involved and their goals" (Franzosi, 2004, pp. 38). The continuing work of Tilly focuses more significantly on balancing the ideographic and nomothetic, the same tension that Singer and Small express in the epigraph to this section. His goal was to create event catalogs but place them within historical and national contexts. As an early example of this work, Tilly, Tilly and Tilly (1975) construct event catalogs of collective violence of at least 20-50 people in France, Germany, and Italy from 1830 to 1960. Their research questions revolved around understanding the nature of contention as a reaction to changing conditions of urbanization and industrialization of Western Europe. Their data sources are varied; in the tradition of social historians, they rely on primary sources and archives, rather than secondary accounts and handbooks (although these too are consulted). Also in that social historian tradition, Tilly focuses much more on tracing narratives of contention rather constructing regression tables or time-series graphs. While the overarching purpose of a work like Tilly, Tilly and Tilly (1975) is to trace the structural conditions underlying contention during European modernization, in the process Tilly and his collaborators also articulate the shifting nature of contention, how those targets of contention have changed over time, and the changing language of articulating those issues. Tilly (1978) outlines the theoretical concerns in political contention, that is, a commitment to understanding group interests, organizations, mobilization, and political opportunity and threat.

Tarrow also describes a body of work called *event histories*. This work focuses less on thick description of historical events, which may include a discussion of movement discourses and narratives from movement participants themselves. Rather it concerns itself with comprehensiveness of event cataloging, conversion of event narratives into quantifiable variables, and use of those variables in statistical analyses. His exemplar is Olzak (1992) and her study of ethnic conflict and protest. Olzak focuses on the changing processes of ethnic competition triggered by economic changes, immigration, the growth of unions, and local labor markets in the United States from 1877 to 1914. Using quantitative event history analysis, she examines the determinants of ethnic conflict and protest in 77 American cities. She uses a mix of the NYTI and historical secondary sources as her basis for event counts. The duration between discrete events is the dependent variable in her analysis. Olzak (1992) has fewer vivid descriptions of the events in their historical context of Tilly's work and involves more discussion of relevant regression machinery, model specification, and time-series plots. There is rarely the description of a single historical event, but much more attention is paid to descriptions of over-time processes of ethnic and labor market competition at the turn of the century. Less attention is paid towards the forms of contention and more attention is given to type of ethnic conflict or protest (minority versus majority, minority versus majority). Similarly, McAdam (1982) relies heavily on event history data to explain the emergence and decline of the Black insurgency and Civil Rights Movement, considering also the surrounding political context and opportunities for the movement. He constructs a dataset of movement-initiated events of the Black insurgency from 1930 to 1970 in the United States. Like Olzak, his source is also the NYTI. While McAdam uses these data for some statistical analysis, such as charting the rise and fall of movement-initiated events, much his analysis is paired with a qualitative histories, public opinion data, and fundraising data. His periodization of the Civil Rights Movement is paired with histories of movement actions and other commentaries. Lastly, Tarrow himself is an exemplar in this tradition with a focus on cycles of collective action in Italy (1989). He is concerned with tracing the arc of social movements and insurgency in Italy from 1965 to 1975, including identifying political opportunity structures, shifting repertoires of action, framing processes of movement agents, the shifting and growing social movement sector, and finally identifying movement demobilization. He relies on one major Italian newspaper and supplements the events drawn from this newspaper with interviews, histories, and other secondary sources. His own narrative uses these data in conjunction with detail and narrative of the interactions of movement actors and the state, not opting for any sophisticated statistical modeling.

Tarrow's work on Italian politics precedes much of the contemporary protest event data projects with a European focus. That tradition began in earnest with the work by Kriesi and colleagues (Kriesi et al., 1992, 1995) and their work on new social movements. This work is a departure from US-based scholars because it relies on cross-national comparisons. It attempts to understand not only temporal variation of movement activity within a single country, but also to compare that movement activity in multiple countries. Kriesi et al. focuses on the Federal Republic of Germany, the Netherlands, France, and Switzerland from 1977 to 1988. The early work on new social movements focuses on political opportunity structures and how different state environments affect levels of mobilization. They coded for forms, issues, organizations, and government reactions. Like Tarrow, Kriesi et al. use a single newspaper for each country. But Kriesi et al.'s methodology is novel insofar as they sample only on Mondays, intended to reduce the labor of coding every newspaper-day but also to capture major protest events occurring on weekends. The other major European project created around this time period focused on Western Germany in the post-war period (Rucht and Ohlemacher, 1992; Rucht and Neidhardt, 1999). The original Prodat project was intended to cover protest from 1949 to the 1989 and the reintegration of East and West Germany. The sampling strategy is similar to the work of Kriesi et al., focusing on Mondays. One criticism of sampling on Mondays, however, is that labor actions (e.g. strikes, walkouts, work stoppages), by definition, take place during the week. As a remedy, in addition to sampling on Mondays, they also sample one full week in each month. Prodat also differs from the Kriesi et al. project by using two newspapers instead of one, an attempt to create a more complete catalog of events.

A separate line of inquiry developed by Franzosi has focused on a generalizable grammar for parsing events writ large. The focus here is a method called *quantitative narrative analysis*. This method looks somewhat similar to the approach taken by Schrodt and colleagues; a news text is analyzed for subject-verb-object triads and networks of actors, targets, and the actions which connect them are created (Franzosi, 2010). This approach has been used to analyze the temporal dynamics of strikes in Italy (Franzosi, 2006).

Much of the work in protest event data has gone the way of the event history approach[2]. The future of protest event data points in two notable directions which improve on this

[2]Hutter (2014*a*) also discusses the body of work in "political claims analysis" (Koopmans and Statham, 1999), a type of research which seeks to combine the quantitative aspects of protest event analysis with qualitative, discursive analysis of claims of movement actors. However, this method has not seen widespread adoption.

framework. First, while no data project is free from its author's research priorities (Hutter, 2014*b*), more researchers have created datasets which are not designed around a specific set of research questions or a specific research project. Instead, they are more focused on generating a general-purpose dataset with longer temporal coverage under a single coding protocol. The Prodat project has already been mentioned; this project has been extended until 2002 under the direction of Dieter Rucht (Rucht and Teune, N.d.). Similarly, the Dynamics of Collective Action (DoCA) project is an extensive effort to track all protest events from 1960 to 1995 from the *New York Times* (NYT) (McAdam et al., N.d.). Rather than relying on NYTI or NYT abstracts, the project looks at the full text of NYT articles. It codes for a range of variables concerning form, initiating groups, issue, target, size, violence, ethnic conflict, and state repression. The original cross-national European data from Kriesi et al. has received a update within the scope of the Political Conflict in Europe in the Shadow of the Great Repression (POLCON) project (Kriesi, N.d.). The original four countries have been expanded to six (adding Britain and Austria). The recent work from this project has focused on the new cleavages between left and right politics in Western Europe (Kriesi et al., 2012; Hutter, 2014*b*), and has continued with the methodology of sampling on Mondays and using a single newspaper per country.

Second, there has been a more recent move towards automating at least part of the coding process. Within the social movements tradition, earlier attempts to help automate this process with technology have failed or not provided the level of detail the researchers were expecting (Imig and Tarrow, 2001). Innovation has come from manipulating search strings in newspaper databases (Maney and Oliver, 2001) and using activist-based web sites as news sources (Almeida and Lichbach, 2003). The fourth edition of the *World Handbook* noted above has focused much more on contentious politics and protest. Relying on a subset of the IDEA ontology and the VRA Reader, they produced a cross-national data set of over 250,000 events (Jenkins et al., N.d.). Lastly, several researchers have focused on a fully automated solution, combining machine learning, named-entity recognition, and other natural language processing and computational linguistics tools (Wueest, Rothenhäusler and Hutter, 2013; Hanna, 2015; Marakov et al., 2015).

1.2 Comparing data branches

Political interaction and protest event data share similar origins and ostensibly could be used to study similar processes. The PANDA protocol and CAMEO ontology mentioned above have several categories for protest event data. Similarly, both DoCA and Prodat extensively document details about the challengers to state power, whether protesters and police enacted violence, and how many people were injured or killed at an event. However, there are a number of reasons that the political interaction / conflict event tradition and the protest event tradition differ significantly. This section outlines the differences in these two traditions by addressing several significant theoretical concerns of the social movements literature, and explaining how those concerns are not adequately expressed in conflict data.

In general, political interaction data is oriented for cross-national, large-N research on state stability, and conflict processes and outcomes. Accordingly, their event categories tend to be comparatively sparse. The paradigm of political interaction data envisions a Westphalia interstate system of sovereign nation-states, some which are more stable than others. Even with the more recent focus on pinpointing the location of events to a substate level (e.g. ACLED and ICEWS contain data on latitude and longitude) and protest (e.g. PANDA and SPEED), these datasets and the scholars who created them are still concerned with nation-state stability and conflict processes, rather than on mobilization processes and outcomes of contention. For instance, the SPEED project, possibly the project with the most descriptive focus on protest in this tradition, holds fast to this view. In a white paper entitled "Definitions of Destabilizing Events," the authors offer an explanation of collective action reminiscent of Smelser's early work:

> Societies are composed of a complex network of equilibria that reflect pre-

> vailing behavioral norms and practices. These equilibria exist in every sector of society (social, economic, political, cultural, etc.). Without them societies could not function smoothly and efficiently; ... Changes in societal equilibria can happen in a variety of ways; they can be incremental, revolutionary, consensual, conflictual, violent, peaceful, etc. But these changes seldom happen easily. Even in "healthy" societies (i.e., those with effective mechanisms for the articulation of demands and the resolution of conflicts) changes in societal equilibria are difficult. (Cline Center for Democracy, 2013, pp. 3)

Protest event data, on the other hand, takes its cues from the establishing documents of the subfield, especially those stemming from the research agendas of Tilly, Tarrow, Olzak, Kriesi, Rucht, and McAdam. Because of the establishment of this tradition, protest event data is rarely dyadic in the way conflict data are. The interaction is often between one or more challengers and the state, but the idea that multiple groups are interacting in important ways that may or may not involve the state has existed from the beginning (Tilly, 1978, esp. Ch. 3).

To understand more deeply how political interaction and conflict data diverge from social movement scholarship priorities, I focus on four main variables: *form*, *claims/issues*, *initiating groups and organizations*, and lastly *targets*.

Protest Forms and Claims: Event Ontologies and Repertoires of Contention

Collecting information on what social movement actors are actually doing at a protest event is at the crux of protest event data. It is tied to the basic definitions of what qualifies as a protest and contentious activity. The idea of *repertoires of contentions* has been at the root of movement scholarship since its inception (e.g. Tilly, 1978; Tarrow, 1989). Tilly notes how repertoires reflect the what are readily "available" to a group, whether that entails the cultural scripts which are suitable, the opportunities which allow for a particular action, or the resources which are at the disposal of movement activists to carry out actions (1978, pp. 151-159).

Accordingly, movement scholarship has focused on the conditions which produce shifts in the repertoires of contention. Tilly has focused on how repertoires have changed with large structural processes of modernization in Western Europe (Tilly, Tilly and Tilly, 1975; Tilly, 1986, 1995). Tarrow (1989) studies how repertoires change within a single cycle of contention in Italy, while Beissinger (2002) explores repertoire shift in the breakdown of the former Soviet Union. The focus on resource mobilization and the political process model led McAdam (1982) to explore the changes to repertoires as conditions for mobilization changed within different stages of Black insurgency, and how protest tactics evolved when movement opponents (white supremacists and state policing forces) found methods of counteracting them (McAdam, 1983). Forms of protest also change as new technological tools become available (Hanna, 2013; Earl and Kimport, 2011), as thresholds of participants are met and more potential participants decide to join (Lohmann, 1994), and new cultural scripts of action are evoked by movement leaders (Moaddel, 1992).

Understanding the concept of repertoires of contention and how they change over time makes it clear that the concept of protest form is significantly different from the projects in conflict data tradition. Many of the existing ontologies fail to capture more than a limited set of forms (e.g. violent versus non-violent, legal versus illegal). Within the most recent CAMEO ontology (Schrodt, 2012*a*, pp. 66–72), there are 26 different events which fall under the rubric of a protest event. These can be distilled into five forms and four claims (plus another "unspecified" category for each). Figure 1.1a lists all the forms available in this ontology, which include "demonstrate or rally", "commit hunger strike", "conduct strike or boycott", "obstruct passage", and "protest violently or riot". This list is very sparse compared to Tilly's examples of historical actions: "[h]ijacking, mutiny, machine breaking, charivaris [a crowd gathering to mock a remarried widower], village fights, tax rebellions, foo[d] riots, collective self-immolation, lynching, vendetta" (Tilly, 1978, pp. 153). This limitation of forms, of course, reflects a desire for abstraction, the abilities for distinguishing between different types of events by the machine coded software, and different research

Code	Form
140	Engage in political dissent
141*	Demonstrate or rally
142*	Conduct hunger strike
143*	Conduct strike or boycott
144*	Obstruct passage
145*	Protest violently or riot

(a) Forms

Code	Issue
14*1	Leadership change
14*2	Policy change
14*3	Rights
14*4	Change in institutions, regime
-	Not specified

(b) Claims

Figure 1.1: Protest events in CAMEO ontology, from Schrodt (2012*a*, pp. 66–72)

priorities. In terms of the latter, rarely are conflict forecasters or foreign policy experts concerned in the difference of political outcomes when villagers engage in a tax rebellion versus breaking of machines. They are typically focused on nation-state instability or propensities towards civil conflict. Therefore the distinction tends to be much simpler: is it a labor action? Is it illegal? Is it violent? And so on. However, the lack of detail to the form in these datasets make them ill-suited for many types of protest analyses.

Second, movement scholarship has concerned itself with a wide variety of *claims*. The number of movement claims are as numerous as movements themselves, but enduring themes surround ethnic and racial rights, abortion and women's health, environmentalism, labor and work, social service provision, and general political malfeasance and ill-provision towards the citizens of a polity. As Tilly (1978, pp. 59-60) notes, classical explanations of mobilization – such as those of Marx, Weber, or Durkheim – tend to take demands for granted, or ground them in large structural processes, such as class or group conflict, or transitions from types of social solidarity. Political interaction scholarship would seem to make similar assumptions. Movement scholars tend to be interested in understanding the claims of movement actors and the mobilizing frames which they construct around them (Benford and Snow, 2000). They want to understand how frames get constructed (Gamson and Modigliani, 1989) and how they are part of the face-to-face mobilizing process (Snow et al., 1986). In short, movement scholars take seriously the constructionist nature of movement demands and how this is part of, not separate to, the act of mobilization.

The status of political claims in movement scholarship differs greatly from how political interaction data treats them. Figure 1.1b notes the different claims available within the CAMEO ontology: “leadership change”, “policy change”, “rights”, “change in institutions, regime”. These do not tell us much about the content of the claims of movement actors. We may be able infer something about “rights” if the protest actor is identified as a member of some minority group. Otherwise, these data do not provide for any type of texturing in understanding the demands of movement participants. In addition to being few in number, they are more or less conflated with the issue of protest target. But the actor ontology part of CAMEO addresses target more fully, as well as groups which are involved with protest.

Targets and Movement Groups: Actor Ontologies and Initiating Groups

So far I have focused only on the character of the event itself and the claims of the protesters. Within the language of political interaction data, this focus is on the events part of the ontology. However, these ontologies contain a separate component focused on interaction: *actors* and their *targets*. The actor-target dyad is the typical unit of analysis for these systems. Thus, what movement scholars typically label as initiating groups, social movement organizations, and targets can be drawn from the actor ontologies.

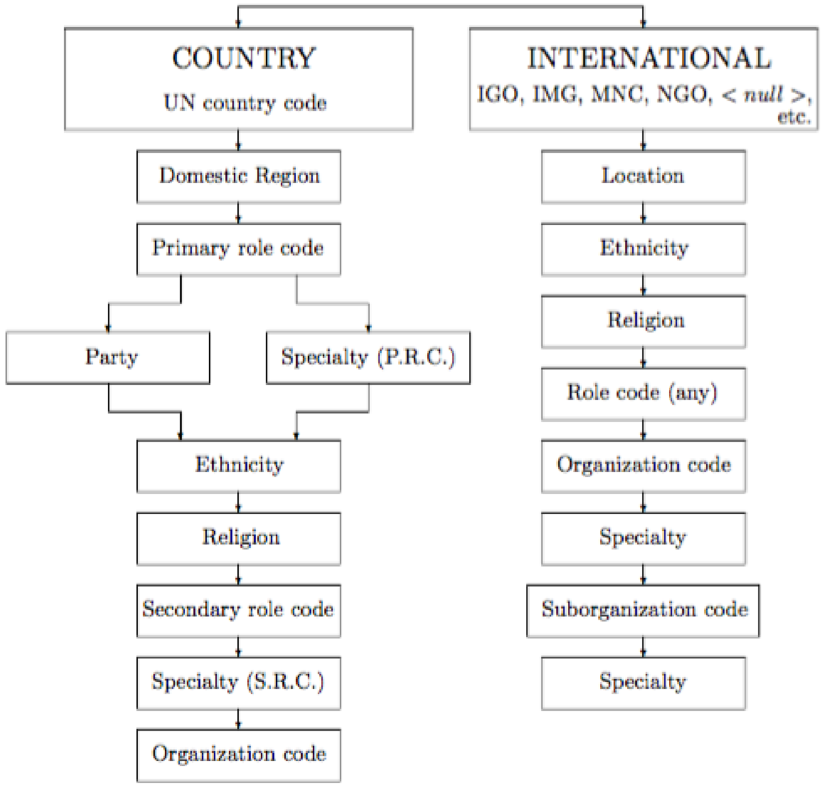


Figure 1.2: Actor ontology coding hierarchy in CAMEO, from Schrodt (2012*a*, pp. 90)

Before describing the differences in targets, initiating groups, and organizations between the two lineages, it is worthwhile to delve into the method of determining a specific actor for CAMEO. Actors in CAMEO represent several different types of entities: individuals, political parties, governments and their constitutive departments and appendages, rebel militias, ethnic and religious groups, and, most relevant for our purposes, opposition movements, organizations, and leaders. This flexibility allows for an almost unlimited number of different actor combinations to be encoded. Figure 1.2 displays the algorithm of determining the relevant actor code in the CAMEO actor ontology. The actor is first delineated by its territory of operation, whether that is a state, a non-state region, or on an international level. As an example, say we were to describe the NAACP. The initial code to denote territory would be `USA`. The actor is then classified into a certain role. This is often called "sector" in the political interaction data literature. These roles include governments, labor organizations, and political parties. In the case of the NAACP, the role would be `OPP`, that is, "political opposition: opposition parties, individuals, and anti-government activists" (Schrodt, 2012*a*, pp. 93, table 3.1). After that, the actor is optionally characterized by ethnicity and religion. We could characterize our current exemplar as `afa`, for "Black-African". It probably would not be an extreme assumption to guess that the attempt to construct mutually exclusive ethnicity categories in actor ontology does not allow for a more specific ethnic category such as "African-American." The ontology, however, does allow for "Native American" (`nai`). All together, then, the NAACP could be coded as `USAOPPafa`. Lastly, particular organizations may be assigned their own organization code if they show up often enough in news reports. An explicit social movement organization reference in the CAMEO codebook is Greenpeace, which is characterized as a difficult case for its geographical scope: "although it is typically thought to be an NGO, it actually functions more as a loose and informal movement with some more formal organizations, such as the Greenpeace Foundation and Greenpeace USA, associated with it." (Schrodt, 2012*a*, pp. 95). Accordingly, Greenpeace obtains its own code `GRP`, so the actor code in its entirety is `NGMENVGRP`, that is Non-governmental Movement (`NGM`) + Environmental (`ENV`) + Greenpeace (`GRP`).

This grammar of actors is surprisingly expressive. It is one area in which it may be more expressive than how movement scholarship typically specifies actors, especially targets. The status of the protest target has traditionally been the state or some state institution. However, we know that within the past 50 years there have been many different targets which are implicated instead of and sometimes along with the state. Soule (2009*a*) notes, for instance, how movement pressure has now moved to multinational corporations and resulted in a shifting of the nature of mobilization to include external and internal pressure on these institutions. Furthermore, the rise of Intergovernmental Organizations (IGOs), such as the World Bank, IMF, and WTO, has added highlighted the role of IGOs in governance, and political and human rights, and therefore made them a prime target for movement activists. Bennett (2003), for instance, notes how digital media has enabled transnational activism, especially against the World Trade Organization. The CAMEO ontology explicit provides for each of these, for instance, multi-national corporations (`MNC`) and the World Trade Organization (`IGOWTO`).

While this ontology is well-suited for protest targets, movement scholarship is also often interested in which social and political groups have initiated protest activity, and whether this group has been formally constituted into a movement collectives. These may be networks of loosely affiliated individuals (Tilly, 1978, pp. 63), ethnic groups (Olzak, 1992), or formal organizations occupying locations in a large social movement field (McCarthy and Zald, 1977; Zald and McCarthy, 1979). The CAMEO ontology does accommodate for certain interest- or ethnically-based organizations. Students uses the role code `EDU`, while labor is `LAB` and generic "human rights actors" is `HRI`.

However, we run into a stumbling block when it comes to organizations. A predefined actor ontology requires frequent updating. If there is to be any reference to explicit organizations, then it must be baked into the ontology. For the type of organization characteristic of political interaction data, this rate of change may be acceptable. Political interaction projects most often formulate actors and targets either as states and their constituent organs or formally-constituted non-state challengers (e.g. rebel groups, opposition parties and their paramilitary arms, militias). However, this is untenable for tracking social movement organizations. Movement organizations and groups are often networks of loosely affiliated individuals, temporary alliances and coalitions, or groups which last for only for several months or a few years. For instance, Wang and Soule note that with after data cleaning, they obtain 4,814 unique organizations in the DoCA dataset (2012, Appendix A). The actor ontology approach, thus, enshrines particular assumptions about the character of movement organizations and state challengers, which is unsuitable for a movement-focused approach.

1.3 Comparing datasets

While the previous section compared the two data lineages in light of the priorities of social movement scholarship, in this section I endeavor to concretize these comparisons by looking at two datasets: the Dynamics of Collective Action (DoCA) project and the Social, Political, and Economic Event Database (SPEED) project. As mentioned above, DoCA attempts to track all protest events from 1960 to 1995 from the *NYT* (McAdam et al., N.d.). It is squarely in the protest event data tradition, guided by prominent social movements researchers (Doug McAdam, John McCarthy, Susan Olzak, and Sarah Soule). I use DoCA because it is considered the state-of-the-art in protest event data for the United States, and it or one of its predecessors has been used in over a dozen studies in the past 15 years[3]. The SPEED project, on the other hand, attempts to provide data on several different types of action: political violence, mass expression (including protest), and disruptive state acts (Nardulli, Althaus and Hayes, 2015). Its latest release includes events from 1946 to 2005 extracted from the *NYT*. SPEED is unique for datasets in the conflict data tradition given that it pays a significant amount of analytical attention to the mass expression event and tracks many of the relevant variables more comprehensively than other conflict data systems. I choose SPEED for three reasons: for one, it has the most temporal and spatial overlap with DoCA and thus is ripe for comparison. Similar datasets which use a KEDS-like system only begin in the 1990s (e.g. King and Lowe, 2008), and those which have prior coverage only cover the Middle East (e.g. Schrodt, 2015*b*). Second, it is coded from the same news source as DoCA, the *New York Times*. Therefore, comparative differences are not a result of newspaper selection bias, but rather the bias introduced by the coding protocol and data pipeline decisions. Third, as far as conflict data goes, SPEED gives a special attention to protest and nonviolent political contention, and provides many more details with regard to those events. This makes it a type of "best case" test for datasets working

[3] `http://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/node/13` lists the current studies which have used the dataset.

within that lineage. It is still clearly within that lineage, however, given to the status it accords to protest events, as noted by its language around "destabilizing events" quoted above. SPEED itself has only recently had a public release, and to date has only been used in one study related to climate change and state instability (Nardulli, Peyton and Bajjalieh, 2015). But we can expect to see its use in future publications.

I use a landmark work in social movement literature as a comparison for the two datasets, namely Doug McAdam's *Political Process and the Development of the Black Insurgency, 1930-1970* (1982). This book represents the first articulation of the political process model of social movement mobilization. In a word, the model questions the status that the resource mobilization perspective grants to the significance and impact of outside political support for deprived groups. Instead, the political process model examines how indigenous organizations are central to the process of gaining the support of powerful political actors. That, along with socioeconomic changes, expanding political opportunities, and the cultural schemata of insurgency leads to social movement mobilization (McAdam, 1982, Ch. 3). The political process model is considered canon in social movement theory, and, although it is not without its critics (e.g. Goodwin and Jasper, 1999), it is represents a central theory for movement scholars.

In this comparison, I focus on two periods within the Black insurgency covered in McAdam's work. Namely, I am interested in the dynamics of movement activity during the heyday of Black insurgency (1961-1965) and the subsequent decline of the movement (1966-1970). In each period, McAdam is concerned with several different facets of movement activity. I will enumerate each of those below and note their significance to McAdam's model.

Hypotheses

In the first instance, we would expect the basic rate of movement events to be roughly the same. Ostensibly, the two datasets are tracking the same events from the same news source.

Within the period of 1960 to 1970, the Civil Rights Movement was at its most active. It is in 1960 that students in Greensboro, North Carolina pioneer the sit-in at the Woolworth's lunch counter; this tactic then diffuses to other segregated institutions and campuses (McAdam, 1983; Andrews and Biggs, 2006). Later on in this period, movement activists introduce the jail-in, freedom rides, city-wide civil rights campaigns, and bus boycotts. Figure 1.3 plots the frequency of these events from 1955 to 1971. The largest peaks are the coordinated movement campaigns in Birmingham (April-May 1963) and Selma (January-March 1965). The second largest peak coincides with the sit-ins in February 1960. McAdam characterizes describes the strategic choices of the Birmingham and Selma campaigns by the Southern Christian Leadership Conference (SCLC) and Martin Luther King Jr., intentionally chosen to evoke a violent response by white supremacists and state actors. Quoting Howard (1968), he notes: "King's Birmingham innovation was pre-eminently strategic. Its essence was not merely more refined tactics, but the selection of a target city which had as its Commissioner of Public Safety 'Bull' Connor, a notorious racist and hothead who could be depended on *not* to respond nonviolently" (1982, pp. 178). Similarly, Selma is also chosen to evoke white violence:

> ...on March 9, state troopers attacked and brutally beat some 525 persons who were attempting to begin a protest march to Montgomery. Later that same day, the Reverend James Reeb, a march participant, was beaten to death by a group of whites. Finally, on March 25, following the triumphal completion of the twice interrupted Selma-to-Montgomery march, a white volunteer, Mrs. Viola Liuzzo, was shot and killed while transporting the marchers back to Selma from the state capital (1982, pp. 179)

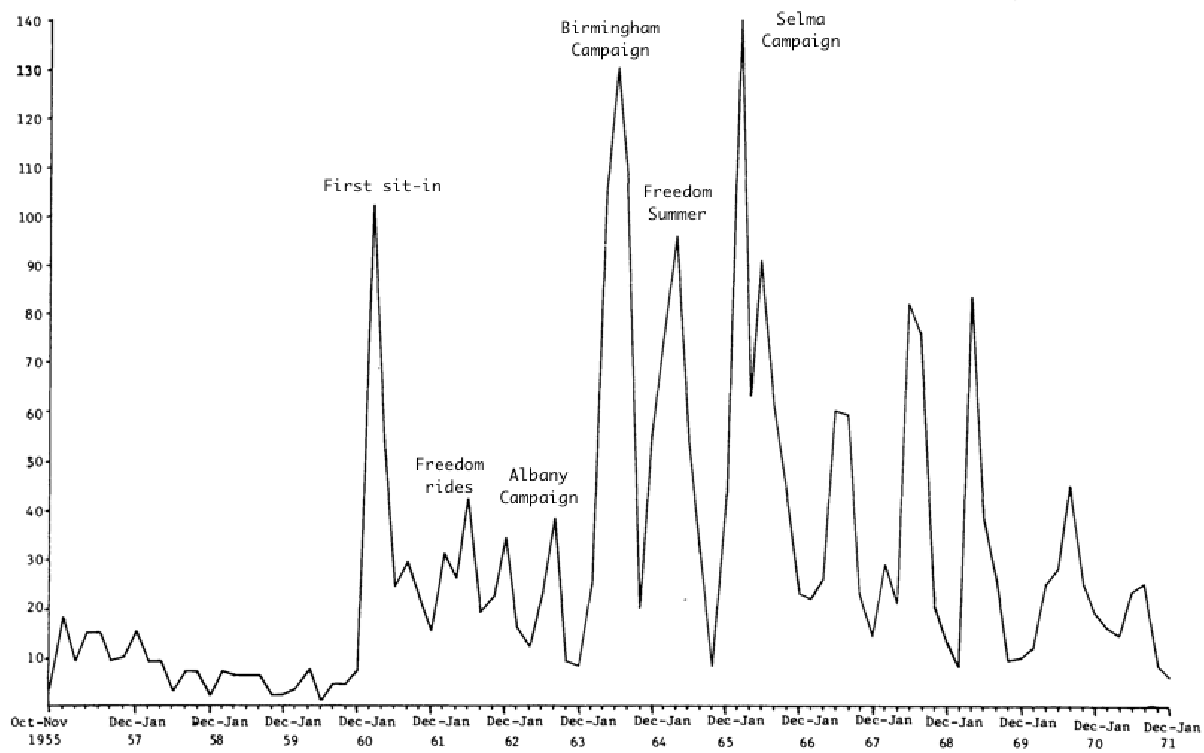


Figure 1. Movement-Initiated Actions, Oct–Nov 1955 through Dec–Jan 1971

Source: *Annual Index* of the *New York Times*, 1955–1971

Figure 1.3: Civil Rights Movement actions 1955-1971. Graph from McAdam (1983, pp. 739), movement labels from (McAdam, 1982, pp. 165)

The problem of measuring the simple frequency of events is compounded, however, by our ability to measure the number of events which are initiated by the Black insurgency. That is, in both datasets, we need to have the ability to discern the initiating groups of each event. The CAMEO ontology notes this by identifying the "sector" of the actor or an ethnic or religious group. Recall that the code `USAOPPafa` could be used to retrieve Black opposition organizations. Although SPEED's own protocol is somewhat different and will be discussed below, given that the tradition of conflict data is attuned to these kind of specificities, we can expect it to track the dynamics of movement-initiated events. Therefore, this leads to my first three hypotheses.

> *Hypothesis 1a (H1a)*: We will see similar dynamics between DoCA and SPEED in the frequency of all events from 1960 to 1970.

H1b: DoCA and SPEED will be able to discern between movement-initiated and non-movement initiated events.

H1c: We will see similar dynamics between DoCA and SPEED in the frequency of movement-initiated events.

Second, one of the main theoretical points of the McAdam's book is how prior indigenous organizations are fundamental to the mobilization process, and that once established, movements must establish their own organizations which will take over the primary activity of movement-building. McAdam characterizes groups within the Black church and the NAACP as indigenous organizations, while organizations which emerge organically from the movement itself are the SCLC, Congress of Racial Equality (CORE), and the Student Nonviolent Coordinating Committee (SNCC). Together, these are the "Big Four" organizations of the Civil Rights Movement. A second major claim is that movement organizations were able to concentrate their power to initiate the majority of events during the heyday of the movement (1961-65). However, after this period, movement organizations are not able to initiate most of the events, although the NAACP remained a major event initiator.

In political interaction data, the organizations which receive standing are those which enjoy a particular standing and have gained recognition through the state or state actors. As noted above, these are often opposition parties or large, international movement organizations (e.g. Greenpeace, Amnesty International). Within the CAMEO ontology, there is not a method of specifying an organization which does not have the standing of one of these organizations. SPEED is different from CAMEO because it does not limit organizations which are not in the dictionary from entering the dataset. However, given that it is within the conflict data tradition, organizations like the SCLC or CORE may not interact significantly enough with the state or act enough like a state challenger to enter into the analysis. This leads to the next three hypotheses.

H2a: DocA will trace the shift in organizational participation from 1961-65 to

1966-70.

H2b: SPEED will not report on the Big Four organizations.

H2c: DoCA will highlight this shift more accurately than SPEED.

Third, McAdam argues that these new movement organizations were able to organize effectively by concentrating their activity into a limited set of geographic locations. This geographic consolidation was primarily centered on campaigns in several cities and states in the Deep South. This includes the Mississippi Freedom Rides, and the Birmingham, Selma, and Albany (Georgia) community-based campaigns. These choices were closely tied both to strategic tactics of confrontation with White violence and state actors, and to building networks with indigenous organizations. The choice also offset resource disadvantages by centering in a few contentious areas. However, in the demobilization period, the growing diffusion of movement events which moved away from the Deep South and into the Mid-Atlantic and the North represented a shift from strategic movement action and proved to sever existing ties with indigenous organizations. Furthermore, the success of creating an indigenous organization network outside of the South proved difficult for SCLC and other organizations, leaving them at the mercy of external support (1982, pp. 190-91).

Given that both protest event data and recent political interaction data are both concerned with specificity of location, we would expect that both of the datasets would be able to account for the dynamic of geographic diffusion away from the Deep South in the 1966-1970 demobilization period. SPEED provides latitude and longitude for each event, while DoCA specifies at least one city and state for an event.

H3a: DoCA and SPEED will be able to trace changes in geographical concentration.

H3b: DoCA and SPEED will both trace this change as predicted.

Fourth, the variety of claims articulated in the movement literature differs widely from those in the conflict data tradition. Within the Civil Rights Movement, McAdam notes that during the heyday of the movement, there was a singular focus around integration and desegregation, especially of public services and transportation but also in housing and education (1982, pp. 153). However, the situation changes significantly during the latter part of the 1960s, with increasing dissensus between different organizations and factions, and no singular issue emerging as priority (1982, pp. 186). Claims of the movement were more varied, including raising the level of Black political power, anti-police brutality, and Black Pride and independence.

The available claims in both DoCA and SPEED codebooks highlight one of the more major challenges in constructing a generalized dataset for contentious politics. As mentioned earlier, claims by movements can be as varied as movements themselves. One of the benefits of constructing a coding system for a single movement case is the ability to differentiate between the nuanced variety of movement claims and internal contention around these claims.

As noted above, the status of movement claim differs significant between the social movement tradition and the political interaction / conflict tradition. This presents a problem in trying to compare claims in each dataset. The analytical status of movement claims is different within each tradition. The DoCA codebook provides its definition of movement claims:

> **Claims Seek to Change System**. A collective action event contains a claim indicating that the group seeks to change the society in some fundamental way; i.e., seeks to redistributed resources, gain new political rights, change (or prevent) some public policy or law, affect public opinion, or fundamentally affect the attitudes, institutions, and/or social culture of civil society. (The Dynamics of Collective Action, N.d., pp. 1)

In this definition, a few things are noticeable: the claim is an intentional and deliberate demand made by the actor. It is against some social structure and the movement actor is cognizant of that. There are some two dozen general claims code in DoCA and 170 specific claim categories.

The status of the movement claim by SPEED looks dramatically different. In their language, the claim is the "origin" of the event. The SPEED documentation does not define origins outright, but suggests the importance of collecting origin variables:

> [M]ost [destabilizing events] are rooted in *something* ... [T]he documentation of inter-temporal changes and cross-sectional differences leads to queries about the factors that generate them – as well as the dynamics of civil unrest over time and across locales. Data on event origins can make important contributions to answering these questions; they can provide the capacity to decompose civil unrest into distinct behavioral domains that are populated by actors with different concerns and time horizons – and who are motivated by different incentives. (Cline Center for Democracy, 2010, pp. 1)

In this quotation, the status of origin is analytically very different than that of the claim. There is no intentionality or agency on the part of the movement actor. Origins cause actors to take specific protest actions. More primacy is given to structural conditions. Actors are the acted upon by structural forces, rather than actors themselves making the claims. Furthermore, SPEED suggests that their levels of aggregation were necessary given the difficulty in accounting for precision in actor demands.

> ...[I]nformation that a mass demonstration or a sit-in is protesting government policies supports an inference that it stems from anti-government sentiments; but it is often difficult to capture much more information on the precise nature of those sentiments (corruption, repression, economic policies, etc.) (Cline Center for Democracy, 2010, pp. 2)

DoCA, however, also using the NYT, attempts to do just this and constructs a comprehensive list of claims. It is not for want of parsimony of the SPEED creators, given that other variables (e.g. state actions) have dozens of possible values. It may be the case that when identifying events cross-nationally, the *NYT* writes about non-US movement actors with much less precision with regard to their claims. But more importantly, the different research priorities of the political interaction tradition make the interrogation of claims less central to their analysis. In the end, SPEED notes only seven origin types in its dataset:

socio-cultural animosities, anti-government sentiments, desire for political rights, desire to retain political power, class-based conflict, desire for personal security, and ecological resource scarcities.

In short, I expect protest event data to be much more focused on the claims of the actors and how these claims shift with changes in organization, in power, and with the cultural frames available at the time. I do not expect that political interaction data to be as nuanced in its interpretation or coding of movement claims.

H4a: DoCA will be able to trace the change in claims.

H4b: DoCa will highlight this shift more accurately than SPEED.

Fifth, the repertoire of contention in the Civil Rights Movement evolved from the heyday to the demobilization period. Tactics shifted as movement actors innovated in response to media environments, public opinion, and the response of the counter-movement (McAdam, 1983; Andrews and Biggs, 2006). Given the focus on repertoires of contention, we would expect that protest event data is expressive enough to capture the over-time variation in tactics, and to enumerate different protest tactics with a good deal of specificity. However, as the review of the political interaction lineage noted, the type of actions accounted for within SPEED will prove lacking.

H5a: DoCA will trace the change in repertoires of contention.

H5b: DoCa will highlight this shift more accurately than SPEED.

Lastly, estimating movement intensity is not simply a question of event frequency. In the case of movements, intensity is also generally judged by size and, less often, duration (Tilly, 1978, pp. 93-97). Many different smaller events do not have the same impact of continually large and escalating events. The shift from sit-ins at lunch counters to large rallies and civic actions represents a major change in the level of mobilization and the strength of the insurgency (Tilly and Rule, 1965; Biggs, 2016). Conflict data also constructs measures of intensity, whether it is on the conflict-cooperation scale, or in building an index to measure event severity (such as COW's nation-months). Accordingly, we would expect similarity in the intensity of events, as expressed by their size.

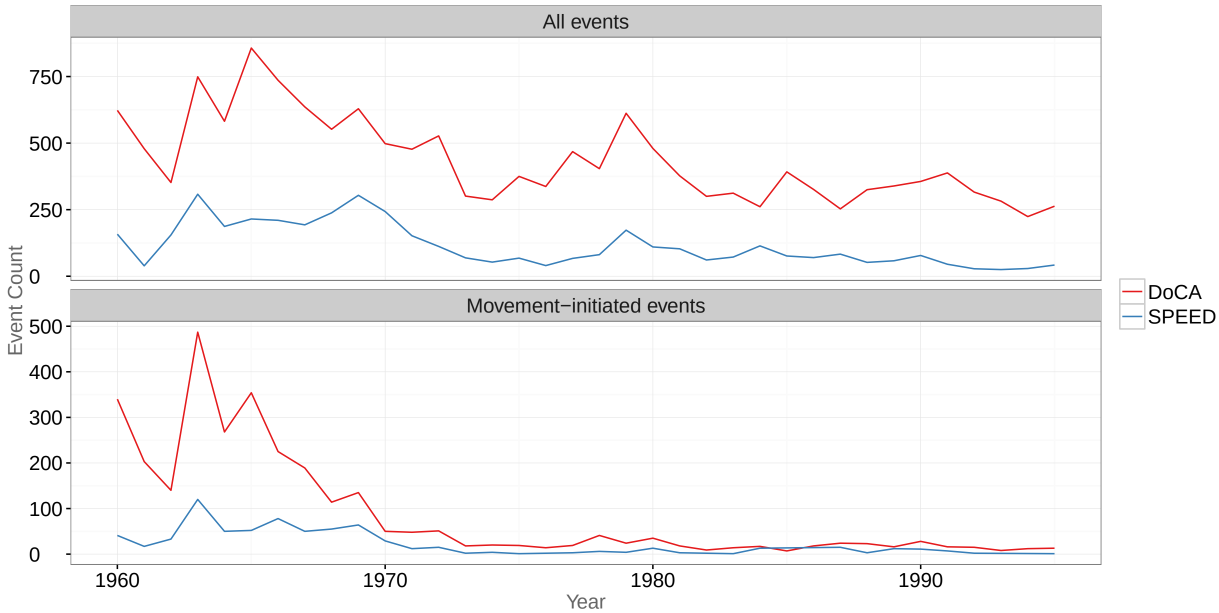
H6: DoCA and SPEED will articulate dynamics of event size in a similar manner.

1.4 Results

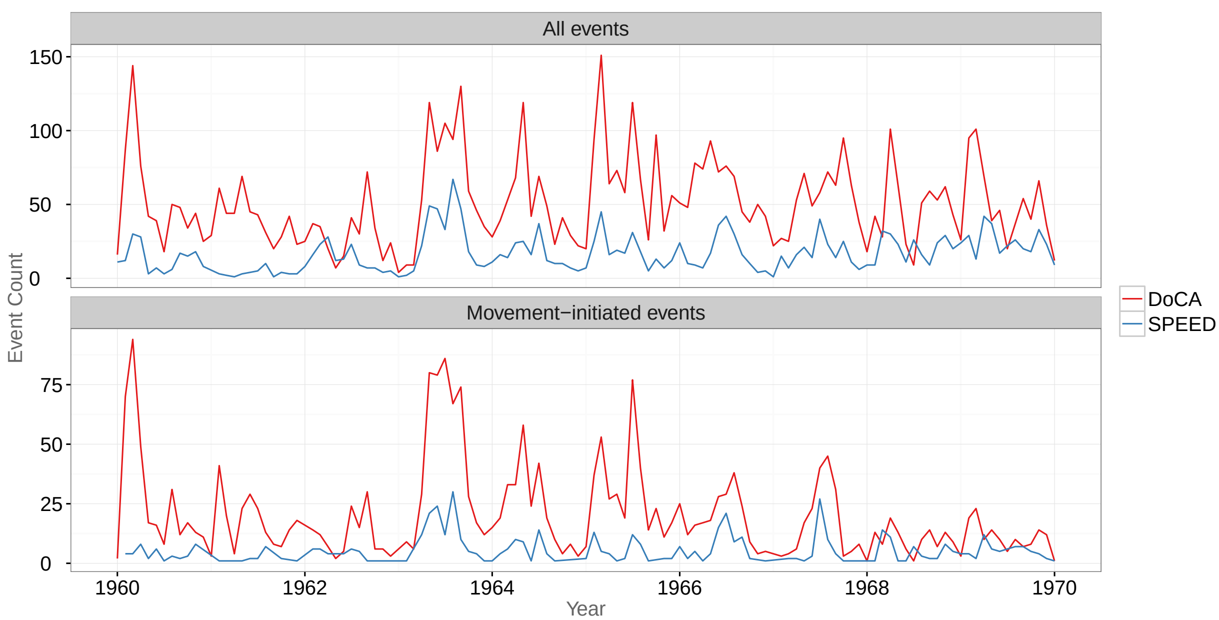
I focus first on the changes in the rates and levels of protests throughout the time period in which we have data from both datasets, 1960 to 1995. The first task entailed identifying movement-initiated events. Within DoCA, African-Americans constituted one of the main initiating group categories. Within SPEED, I relied on a variable related to socio-cultural group initiator. For both datasets, I also included any event which involved one of the Big Four movement organizations (NAACP, CORE, SCLC, and SNCC), as well as the Black Panther Party.

Figures 1.4a and 1.4b displays the over-time frequency for both of the datasets, both for all events and events which were strictly movement-initiated. First, in the 1960-95 time period and in the United States, DoCA contains many more events overall, just under 16,000[4]. SPEED, however, only reports 5,450 events with these same parameters. In terms of movement-initiated events, DoCA reports about 3,000, while SPEED reports 600. Despite the discrepancy in sheer volume, the frequency of the events over the longer time period (Figure 1.4a) looks somewhat similar, with peaks in all events and especially movement-initiated events during the 60s. Figure 1.4b displays the monthly frequency of events in both datasets. In this time period, DoCA reports about 6,700 total events and 2,505 movement events (2,165 from 1961-70); SPEED reports over 2,400 total events and 589 movement

[4]In this analysis, I have not included ethnic and racial conflict events, which were coded without meeting the criteria of movement events and were most likely violent attacks of members from one racial group on members or property of another. I also excluded the events which were primarily labeled as lawsuits in the form or activity variable.



(a) 1960-95, by year



(b) 1960-70, by month

Figure 1.4: Event frequency comparisons in DoCA and SPEED

	1960 - 1995		1960 - 1970	
	Monthly	Yearly	Monthly	Yearly
All events	0.62***	0.78***	0.60***	0.45
Movement-initiatied events	0.59***	0.86***	0.54***	0.61*

*p < .05,** p < .01,*** p < .001

Table 1.1: Pearson correlations between DoCA and SPEED.

events (549 from 1961-70). The peaks in Figure 1.4b matches those in Figure 1.3. There is an early peak during the sit-ins in 1960 and a second set of peaks during the Birmingham and Selma campaigns. While all of these campaigns are visible with the DoCA dataset, SPEED does not indicate any movement activity for the sit-ins. Although DoCA tracks the same general trend as McAdam, it also reports a lower number of events overall.

Table 1.1 notes the Pearson correlations between the datasets, aggregated at the monthly and yearly level. Monthly correlations over the whole period are 0.62 and 0.59 for all events and movement-initiated events, respectively. The correlations increase at the yearly level to 0.78 and 0.86. For the 1960s, the monthly correlation is 0.60 and 0.54, and 0.45 (not statistically significant) and 0.61 at the level of yearly aggregation. These correlations are moderately strong, and suggest SPEED is tracking somewhat similar results as DoCA, if not with the same magnitude and descriptiveness. These results provide some support that DoCA and SPEED are tracking the same kinds of events (H1a), but there are some noticeable silences in SPEED, especially with regard to the sit-ins in 1960. They also suggest we are able to discern between all events and movement-initiated events for both datasets (H1b). However, again, it provides weak support that SPEED is tracking movement-initiated events in the same manner of DoCA (H1c).

Turning to organizations, McAdam notes that in the heyday of the Black insurgency, actions were initiated primarily by movement organizations. Power in the 1955-60 period moved from Southern-based chapters of the NAACP, which initiated some 70% of actions, to the nascent movement-focused organizations of CORE, SCLC, and SNCC. However, after

	DoCA				SPEED			
	1961-65		1966-70		1961-65		1966-70	
Organization	N	%	N	%	N	%	N	%
NAACP	205	25.15	91	25.07	9	21.43	5	16.13
CORE	288	35.34	47	12.95	18	42.86	4	12.90
SCLC	63	7.73	38	10.47	2	4.76	1	3.23
SNCC	63	7.73	28	7.71	7	16.67	2	6.45
Other	196	24.05	159	43.80	6	14.29	19	61.29

Table 1.2: DoCA / SPEED organization comparison

1965, the ability of organizations to initiate action became more diffuse, changing from movement organizations to other groups which were not part of the Big Four. The NAACP, however, was able to maintain its level of activity. McAdam notes the decline of the Big Four could be attributed to organizational diffusion and dissent over goals and tactics. The usual ideological delineation puts the NAACP and the SCLC into an integrationist camp, while the SNCC and CORE become more radicalized and come to rely on more violent means of action. The SCLC was able to maintain a high level of activity up until the death of Martin Luther King Jr. in 1968, after which it saw a decline in activity.

We would expect that the data can track these changes in the activity of the Black insurgency. Both DoCA and SPEED allow for the identification of named organizations. DoCA allows for up to four to be identified, while SPEED allows for one. In this analysis, I focused only on the initial organization in the DoCA dataset. I also merged the different spellings of the organizations with regular expressions, given that previous work with DoCA has noted inconsistencies between named organizations (Wang and Soule, 2012, Appendix A).

Table 1.2 notes the results of this analysis. In the first case, DoCA reports high amounts of activity for the Big Four in the first period. The NAACP and CORE are involved in the largest number of events, constituting over 75% all together. The SCLC and SNCC, however, only are involved with around 7% of events. The low number of events is at odds with the result in McAdam (1982, pp. 154), which puts SCLC's percentage closer to 20%. In the latter period, the NAACP maintains a consistent level of activity with the prior period, as does SNCC. We also see a dramatic decline of participation by CORE. However, against this, SCLC actually takes a larger share of events. Therefore hypothesis 2a is partially correct, save for the results we see with the SCLC.

Moving now to SPEED, contrary to H2b, SPEED does indeed contain information on the Big Four organizations. However, the number of events on which it contains any organization information is small: only 73 of 547 (13.3%) events contain organization information, compared to 1,178 of 2,165 (54.4%) in DoCA. However, SPEED does reflect a similar substantive result, with the Big Four responsible for the majority of events in the earlier period, and the NAACP and CORE at the head. Like in DoCA, there is a similar result with the SCLC, which initiates only about 5% of events. Unlike DoCA and McAdam (1982), SNCC is reported to be responsible for nearly 17% events. In the latter period, NAACP and SCLC remain consistent, but CORE sees the same large decline as it did with DoCA. SNCC also sees a similar decline.

Therefore, hypothesis 2c is somewhat incorrect, that is, DoCA and SPEED trace similar patterns, namely the changing concentration of movement activities from the Big Four organizations to other movement organizations. However, they differ in how well they do this. The SCLC seems to be unrepresented in both datasets in the earlier period. It may be the case that McAdam is including mentions of Martin Luther King Jr. as indications of involvement of the SCLC. He does indeed, include a note about King in the table on pp. 183. Other scholars have highlighted the different and special coverage during movement actions that involved King, typically involving more substantive coverage of movement activities (Amenta et al., 2015). Including his activity as part of the SCLC would sharply change the story we can tell about SCLC activity compared to the one produced here.

Focusing now on location, McAdam notes the concentration of actions in the South during the heyday of the movement, especially in the Deep South[5]. As noted above, this

[5]McAdam (1982, pp. 152) defines the Deep South as Alabama, Georgia, Louisiana, Mississippi, and South Carolina. He also includes two other Southern categories: Middle South (Arkansas, Florida, North

	DoCA				SPEED			
	1961-65		1966-70		1961-65		1966-70	
Region	N	%	N	%	N	%	N	%
Border States	99	6.83	57	8.01	18	6.79	8	3.09
Deep South	543	37.47	130	18.26	121	45.66	64	24.71
Middle South	183	12.63	55	7.72	34	12.83	21	8.11
North Central	122	8.42	135	18.96	14	5.28	68	26.25
Northeast	449	30.99	300	42.13	63	23.77	80	30.89
Other South	1	0.07	0	0.00	0	0.00	0	0.00
West	52	3.59	35	4.92	15	5.66	18	6.95

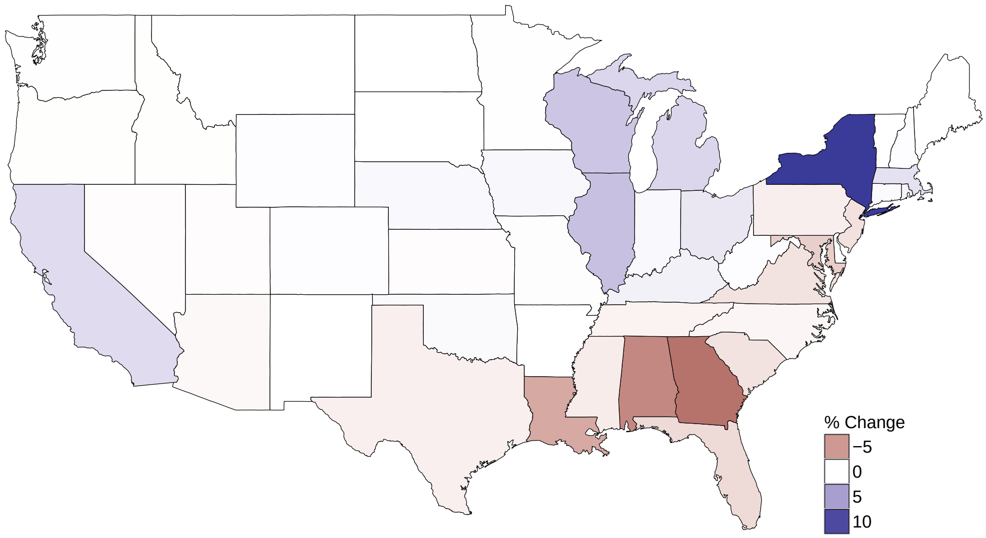
Table 1.3: DoCA / SPEED location comparison

geographic concentration allowed the movement to create networks between indigenous organizations, especially in the Black church and the NAACP. However, during the latter period, actions diffused and entered the North, Middle Atlantic, and West.

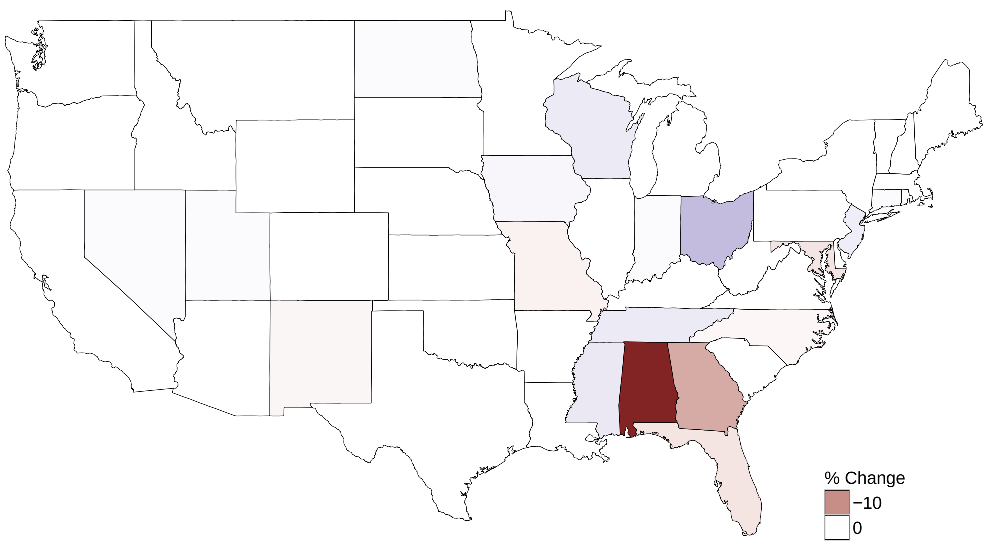
Table 1.3 and Figure 1.5 note the locations of the actions in both periods. In DoCA, 57% of actions occurred in the South. The Deep South itself saw over 37% over actions. Other events took place mainly in the Northeast (31%). In the demobilization period, the number of actions in the South declined to 34% of all actions. The number of events in the Northeast and North Central increased, to 19% and 42%, respectively. Figure 1.5a displays the change in percentages of events in each state. Accordingly, all states in the Deep South report the largest decreases, while New York state saw the largest increase. States in the North Central (Wisconsin, Michigan, Illinois, and Ohio) and the West (California) had modest increases in activity.

As for SPEED, we also see that the majority of events occur in the South in the early period by an even larger margin, about 65%, with the Deep South accounting for 46% of those events. The Northeast also has a large share of events, near 24%. In the latter period, there is also a drop in actions in the South, down to near 36%, 25% in the Deep South. And like DoCA, there are increases in events in the North Central and Northeast. Looking at the

Carolina, Tennessee, Texas, Virginia) and Border States (Kentucky, Maryland, Missouri, Oklahoma, West Virginia, District of Columbia). Delaware is classified as Other South, given that is considered part of the South in Census categories but is not accounted for in the prior lists.



(a) DoCA



(b) SPEED

Figure 1.5: Percent change of movement-initiated events, 1961-65 to 1966-70.

map in Figure 1.5b, the largest drop occurred in Alabama, with Georgia following. Oddly, however, there were very moderate increases in activity in Mississippi and Tennessee, about 2% in each state. The aggregated result in Table 1.3 obscures the slight rise in events in these states. McAdam does not outline the rates of change in individual states, and it may be the case that some states in the South saw a moderate rise in the rate of events.

According to the data presented here, H3a is well-supported. Both DoCA and SPEED are able to track per state changes in events. H3b is mostly supported; both datasets trace the geographical shifts between the heyday and the demobilization period reported by McAdam. However, SPEED has some anomalies which may not be supported by the original analysis.

Moving to claims, McAdam notes that the coherence of claims of the movement allowed it to maintain strength and organizational leadership. However, in the demobilization period, the proliferation of claims and the differences over tactics started to tear the movement apart and result in a diffusion of focus on what the orienting goals of the movement should be.

Claim	1961-65 N	1961-65 %	1966-70 N	1966-70 %
Any Desegregation Claim	645	44.42	50	7.01
African American Civil Rights, general	302	20.80	198	27.77
Anti-Discrimination in Housing or Employment	150	10.33	85	11.92
Anti-Police Brutality/Harassment	93	6.40	86	12.06
Pro-Voting Rights/Political Power	92	6.34	29	4.07
State surveillance/Prosecution of protesters	28	1.93	14	1.96
Education	23	1.58	101	14.17
Black Pride, entrepreneurship, separatist	5	0.34	24	3.37

(a) DoCA

Claim	1961-65 N	1961-65 %	1966-70 N	1966-70 %
Socio-cultural animosities	268	98.53	275	99.64
Anti-government sentiments	141	51.84	111	40.22
Desire for political rights	134	49.26	65	23.55
Desire to retain political power	26	9.56	16	5.80
Class-based conflict	14	5.15	25	9.06
Desire for personal security	3	1.10	1	0.36
Ecological resource scarcities	0	0.00	3	1.09

(b) SPEED

Table 1.4: DoCA / SPEED claim comparison

Table 1.4 notes the changes in claims over the two time periods for DoCA and SPEED[6]. Within DoCA, during the prior movement period, an overwhelming number of events were concerned with desegregation, nearly 44%. This was followed by general African American civil rights claims, anti-discrimination, and anti-police brutality. In the latter period, the focus on desegregation dropped dramatically, to only 7% of claims. General claims rose to nearly 28%, and claims around anti-police brutality rose to 12%. Curiously, education, which had been a mainstay of desegregation claims, rose to 14%. The general trend, however, supports H4a and McAdam's analysis that the demobilization period resulted in disagreement about goals and less of a focus around integration.

The SPEED results are much less conclusive and tell us less about the changing claims

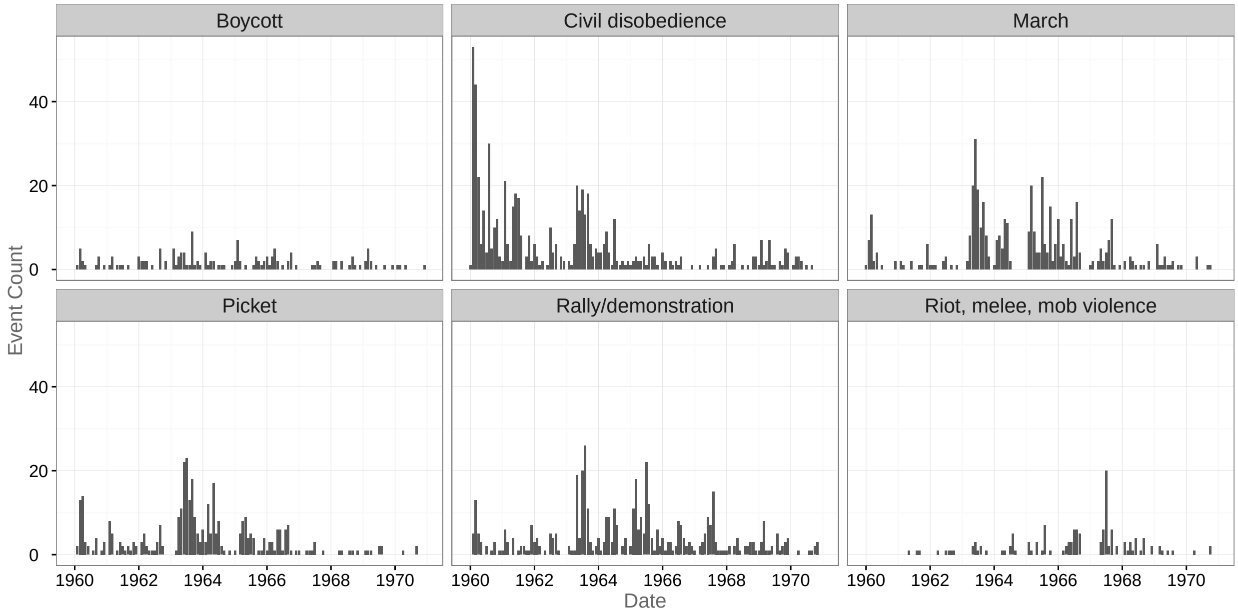
[6]For DoCA, I only included claims associated with African Americans or miscellaneous movement demands. I also filtered out claims which did not appear in over 1% of events in either period.

of movement actors[7]. In Table 1.4b, nearly all events can be associated with socio-cultural animosity over both periods. About half of them are anti-government, 55% in the earlier period and 41% in the latter. Notably, nearly 53% of events in heyday period are coded as a desire for political rights, while 30% are coded as such in the demobilization period. This code may be capturing some of the desegregation claims. Lastly, class-based conflict seems to increase from 6% in the prior period to nearly 13% in the latter. This may indicate the radicalizing, anti-capitalist claims of the movement in this period. However, using the SPEED claims and attempting to map them onto the more expressive set of claims in McAdam's book is an exercise in conjecture. This result provides very good support for H4b.

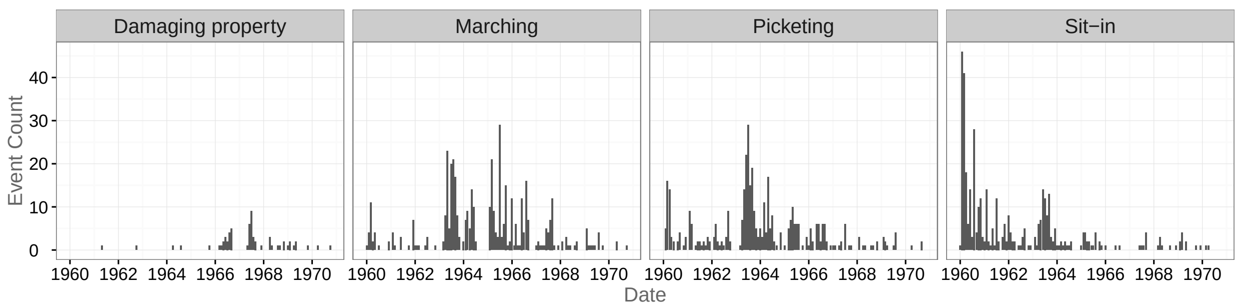
In terms of protest repertoires, McAdam (1982, 1983) documents the interplay between movement actors and counter-movement forces, and how this interplay affected movement tactics. We expect the data to reflect the cycle of protest and its choice of tactic from its repertoire of action. This means sit-ins and civil disobedience in the early 60s, community campaigns in the middle of the decade, and lastly, riots and violent action in the latter part of the cycle.

The form variable is another in which there are some major differences between the dataset grounded in social movements compared to the political interaction / conflict data. The differences between the two are not as drastic as with the claim variable, however. DoCA traces two variables: form and activity. Forms of protest are more general characterizations of the event, while activity are the concrete activities occurring within the event. SPEED has two variables related to protest repertoires: type of political expression and type of political attack. The political expression variable entails several different forms of protest – verbal and broadcast expressions, demonstrations, and strikes. The political attack variable is relevant insofar as it indicates the presence of riots. There are also variables related to types of symbolic expression and advocacy for certain tactics, but these did not contain

[7] I did not filter out any claims in SPEED. However, claims are not mutually exclusive and the percentages in Table 1.4b are out of all events in that time period.



(a) Forms



(b) Activities

Figure 1.6: Protest forms and activities in DoCA, 1960-70

much information about the events under study.

Figure 1.6 presents the results from DoCA. Figure 1.6a shows protest forms across the time period. Under civil disobedience, we see peaks in 1960, which corresponds with the spread of sit-ins during that year. There's a second peak during 1963, which matches the time of the Birmingham campaign. Marches, rallies, and pickets correspond with the community campaigns, although there are continuing spikes in 1967 and 1968. Riots start to rise during the latter part of the decade, which matches McAdam's account. Surprisingly, boycotts do not register as significant throughout, although they are a consistent presence throughout the time period. Moving to the protest activities in Figure 1.6b, sit-ins peak

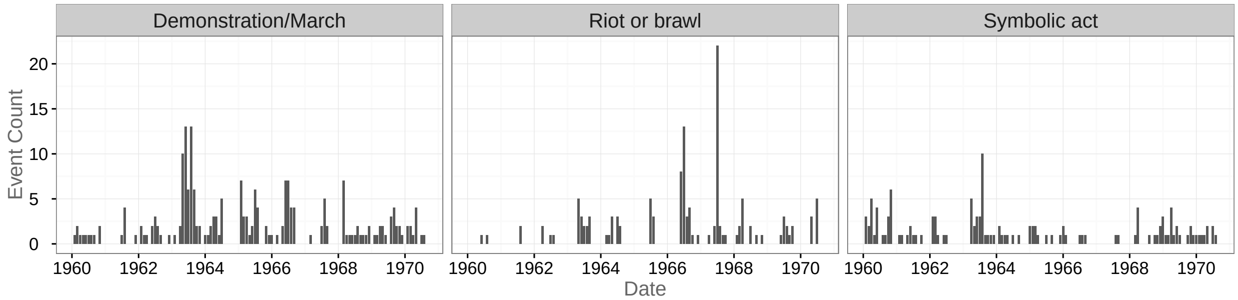
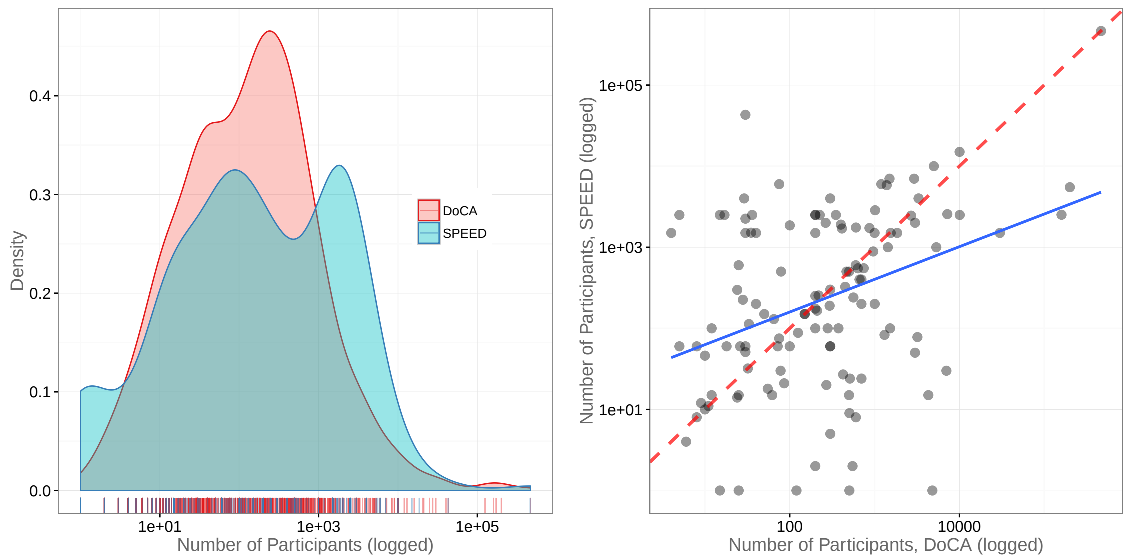


Figure 1.7: Political expression in SPEED

in the 1960s, as expected. Picketing and marching match the community campaigns. Damaging property roughly correlates with the period of rioting. This provides support for H5a, that is, the over-time changes in tactics seem to match the descriptions provided by McAdam (1983).

The results from SPEED are displayed in Figure 1.7. Demonstrations and symbolic acts peak in 1963 – during the Birmingham campaign – and also in 1967 and 1968. Riots also have large peaks during those years, matching prior expectations. However, there seems to be no data showing the large number of sit-ins in 1960. Indeed when we look back at the frequency of movement-initiated events in Figure 1.4b, there is an almost flat line for SPEED during the year of 1960. Comparatively, DoCA has a large spike in events during that year. Lastly, although the "advocacy for tactics" variable in SPEED incorporates civil disobedience, boycotts, and riots, nearly all of the values for this variable are none or missing. This is a curious oversight in this dataset. Overall, however, DoCA matches the dynamics of repertoires of contention described by McAdam. These results provide support for H5b.



(a) (b)

Figure 1.8: Size comparisons of matched days between DoCA and SPEED

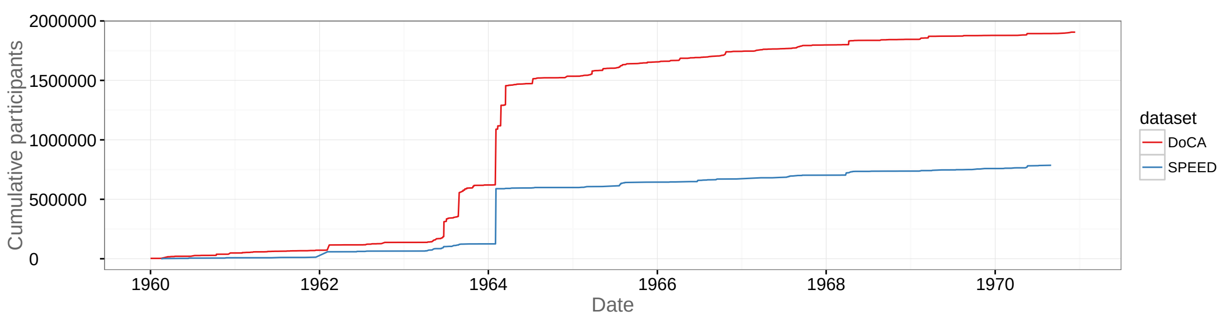


Figure 1.9: DoCA compared to SPEED by cumulative sum of size

Lastly, while McAdam does not focus on size, using event size may tell us more about event dynamics than merely using event frequency. Figures 1.8 and 1.9 display the comparisons between the numbers of participants in SPEED and DoCA. Figure 1.8a is a density/rug plot of the logged number of participants. Most events in DoCA seem to be of similar size and are centered at one peak. The mean number of participants in DoCA is 1872, and the median is 150. However, SPEED has two peaks in event numbers, with one peak having a higher value than the DoCA mean. The SPEED mean is 2869, but the median is also 150. Figure 1.8b displays a scatter plot of the sum of the logged number of participants for events on that day, for all the days in which both datasets report participants (N = 124).

Points which lie on the red line represent values which show agreement. The blue line is the regression line drawn through the points. The slope of the regression line is less than 1, which indicates that, on the whole, DoCA is reporting larger events. However, this result is driven by the fact that DoCA has a several very large events which are not reported in SPEED. That fact, in and of itself, is significant. If SPEED is missing very large events, then there may be some important movement events which it ignores. Lastly, Figure 1.9 displays the over time cumulative sum of daily participant numbers. The two curves have a Pearson correlation of 0.99 (significant at the 0.001 level).

The largest event for both datasets is on February 3, 1964. DoCA reports an article titled "BOYCOTT CRIPPLES CITY SCHOOLS", in which Bayard Rustin and the NAACP organized a boycott of schools by over 464,000 Black and Puerto Rican students in support of full integration of the city's schools[8]. The next largest event in DoCA has the title "200,000 MARCH FOR CIVIL RIGHTS IN ORDERLY WASHINGTON RALLY; PRESIDENT SEES GAIN FOR NEGRO" – the March on Washington on August 28, 1963 in which King gave the "I have a dream" speech[9]. The next three events which have the largest size feature the titles "CHICAGO BOYCOTT DRAWS 172,350"[10], "SCHOOL BOYCOTT IS HALF AS LARGE AS THE FIRST ONE", which was a followup to the February 3 event[11], and "125000 RALLY IN DETROIT TO PROTEST DISCRIMINATION"[12]. They are reported to have involved 172,350, 160,000, and 125,000 participants, respectively. From the information available in SPEED, there does seem to be a report for the March on Washington, but the number of participants is reported to be 2,500. For the other events, there does not appear to be any corresponding event in SPEED.

While the basic contours of the intensity of movement activities seem to correspond between DoCA and SPEED, the latter dataset misses or severely undercounts critical events

[8] `http://www.nytimes.com/1964/02/04/boycott-cripples-city-schools.html`

[9] `http://www.nytimes.com/learning/general/onthisday/big/0828.html`

[10] `http://www.nytimes.com/1964/02/26/chicago-boycott-draws-172350.html`

[11] `http://www.nytimes.com/1964/03/17/school-boycott-is-half-as-large-as-the-first-one.html`

[12] I was unable to find this article in the NYT online archive.

in heyday of the Black insurgency. This does not lend much support to H6.

1.5 Discussion

Overall, DoCA is able to track similar dynamics of the Black insurgency in the 1960s more accurately than SPEED. The frequency of events, the organizations involved, the location of the events, the claims of protesters, and the protest repertoires all seem to mirror the results that are presented by McAdam. SPEED, on the other hand, tracks frequency, location, and organizations well, but fails when it comes to claims and protest repertoires. Moreover, DoCA is able to track large, major events on the Civil Rights Movement, while SPEED does a questionable job of this. Given the focus of conflict data on location and intensity, the failure to provide accurate numbers of size seems like a critical deficiency of this dataset.

Another major limitation of SPEED is the small number of movement-initiated events that are recorded. In the 1960s, SPEED contains only about one-fifth of the number of events compared to DoCA. This may be a consequence of a number of factors, including the number of countries attempts to cover and the machine learning algorithm used to identify relevant events. SPEED contains data on 199 countries for events including protest, political attacks, and state retaliation. Its coverage is meant to account for large, global trends. The majority of SPEED's events, however, are located in the United States (about 10,000 events, or 16%), followed by Israel (4.2%), Lebanon (3.8%), and the UK (3.1%). The incredible focus on the US and Israel seems very much like a result of selection biases of the *NYT*. We know that, for instance, the further away an international newspaper is from the event, the less likely it will be to report on the event (Mueller, 1997). But even with the diffused focus on the events, SPEED coders should have been able to pick up a similar volume of events as DoCA, given that they were drawn from the same source.

A second limitation may be the use of automated methods for filtering out irrelevant events. According to their documentation, SPEED uses a set of machine learning classifiers to filter out irrelevant articles. They use the Naive Bayes classifier for this task, and, after a number of different tests, they report only a 1-4% false negative rate (Leetaru, N.d.). SPEED chooses to optimize to minimize their false negative rate, given that human coders will eventually look at and code the text. It may be the case that they are optimizing for a particular kind of protest event (the mass expression event), therefore the event does not meet the DoCA's criteria for inclusion. This, in turn, would also bias the machine learning classifier's criteria for classification.

Apart from the direct comparison, there are two issues which are not mentioned above which present major differences in the two branches. The first focuses around the problem of selection bias and the types of articles which tend to appear in news sources. The second is the geographic scope of the datasets and how that also may affect the analysis.

The proliferation of different protest event datasets in the 1980s and 1990s started to come under scrutiny by what Hutter (2014*a*) characterizes as a third generation of movement scholars who questioned the representativeness and validity of newspaper data. This led to a deeper interrogation of the biases embedded in newspaper data. The discussion of the various biases of protest event sources has been written on extensively (for thorough reviews, see Earl et al., 2004; Ortiz et al., 2005; Davenport, 2010; Jenkins and Maher, 2016). We know from these studies that newspapers are biased toward larger events, events which are more violent, and events which are closed to the news source, either the newspaper headquarters or a bureau desk in the case of news wires. We also know that in the course of protest reporting, at least in the US, protests are more often than not covered as part of a public order and crime beat, rather than by political or by local interest reporters. Movement scholars have paid a great amount of attention to these issues. Comparatively, conflict scholars have paid much less attention to these issues until very recently (Davenport and Ball, 2002; Weidmann, 2016)[13]. The assumption, echoed by Tilly and Rule (1965), is that collective violence events are assumed to be socially significant enough to be covered

[13]It should also be noted that these scholars are also heavily concerned with protest.

in newspapers. But this is cannot be the starting assumption for areas of the world which have low levels of technological penetration (Weidmann, 2016) and conflict zones in which reporters cannot access or are prevented from reporting (e.g. Syria post-2011 and Boko Haram-controlled areas of Nigeria (Ulfelder, 2015)).

This problem is further exacerbated by the usual cross-national focus by conflict data scholars. Conflict and political interaction data differ drastically from protest event data in the range of countries they cover. These datasets are universally cross-national, focusing on the whole world or a large region of concern (e.g. the Middle East). This makes sense, given the interstate and state-based nature of conflict studies. In contrast, protest event data tends to focus exclusively on a single country or a small group in comparative analyses. The tradition of protest event data is typically associated with scholars studying their own country or a country in which they have subject-matter expertise. Newspapers are coded in their native language. This mimics the more specialized knowledges of sociology, anthropology, area studies, and history, as a sharp distinction from the large-N, cross-national coverage of most conflict data.

The problem of selection bias is then multiplied, possibly by the number of countries involved, and even more so if there is strict reliance on English-language news sources. Some projects – namely the Open Event Data Alliance – are attempting to use many more newspapers and is spearheading a project to construct ontologies in several major languages (including Arabic and Spanish). However, the selection biases inherent in the usual practices of the political interaction tradition are rife and mostly unexamined.

1.6 Conclusion

In this article, I have highlighted the two lineages of political event data in detail by highlighting their histories, their coding processes and data sources, and the variables and format of their final products. This kind of interrogation involves the investigation of many codebooks, methodological appendices, and datasets. Yet as David Singer, the creator of the Correlates of War dataset, notes, “You live and die by your coding rules.”

The lineage of political interaction data is structured around conflict between sovereign nation-states, or serious challenges to the sovereignty of those states by non-institutional challengers, including political parties, armed rebels, and ethnic militias. These data can be both episodic or discrete in nature, and track everything from diplomatic talks to mass killings and genocide. They highlight a fundamental dyadic nature of the events, between two more or less formally constituted entities.

In the other case, the lineage of protest event data focuses less on dyadic relationships between the state and its challengers, but on the repertoires of contention, the claims of movement activists, the initiating groups and organizations involved in the event, and the targets of the protest action. Furthermore, movement scholarship focuses much more on the activity of the movement actors and understands their actions as intentional and based in legitimately constructed grievances.

In comparing these two lineages, it is clear that in the political interaction tradition the conceptualization and operationalization of movement forms, their claims, and social movement organizations is dramatically different and less expressive than those in protest event data analysis. The comparison of DoCA and SPEED highlights that the status of the movement claim has much more to do with movement actors than the structural conditions in which the movement actors are embedded. It also highlights the close attention to the various forms of protest and the dynamics of these forms as the political context – such as responses from counter-movement protesters and from state repression – changes and adapts. Finally, protest event data has given special attention to selection bias and the notable absences which may occur when casting too far of a geographical net.

However, with the development of automated tools for analysis, these two lineages may find some unification yet. Already, sociologists have begun to use some of these data to understand global changes in democratic development and liberalization (Kadivar and

Caren, 2015) and the instability of political forecasts in the face of exogenous shocks in sustained political conflict (Kurzman and Hasnain, 2014). Creating automated systems which account for the priorities of movement scholarship has the potential to significantly change the potential scope and theoretical depth for both social movement and conflict scholars.

2 MPEDS: AUTOMATING THE GENERATION OF PROTEST EVENT DATA

The social media age has drawn vast amounts of attention to modern social movements. Movements such as Black Lives Matter and Occupy Wall Street have reinvigorated discussions about the unequal distribution of income and wealth, the amount of control by multinational corporations and banks, and vast racial disparities in policing, sentencing, and incarceration. As scholarly and public interest in protest increases, there is a growing demand for good data on contentious collective action events in a variety of fields. International relations and foreign policy experts are often interested in using protest event data to forecast political instability and state breakdown. The emerging field of "data journalism" tells narratives about protest activity and political changes around the world. And most relevant for the current project, scholars of social movements and contentious politics need high quality protest event data to understand the emergence, dynamics, and consequences of new social movements and contentious collective action.

However, the lack of high quality protest event data is a chronic issue in social movement research. Comprehensive protest event data with broad spatial and temporal coverage is limited by both the availability of primary sources and speed at which we can code these sources for relevant features for scholarly and practical work. Social scientists have relied primarily on newspapers to gather information about protest events. The biases in using newspapers as primary sources are well-documented by social movement scholars (e.g. Franzosi, 1987; Earl et al., 2004; Ortiz et al., 2005). Biases induced by selective coverage are difficult to address, but incorporating multiple media sources may be an adequate, albeit not perfect, corrective. Given the explosion of electronic archives of newspapers and the availability of new digital media, the potential for identifying protest events is enormous.

With this increasing availability of digital sources from which we can identify protest events, the challenge is to code these sources for collect relevant information. Hand coding newspapers has been the traditional strategy for identifying protest events within social movements scholarship (Hutter (2014*a*) provides a recent review of many protest event data and analysis projects). The advantage of this approach is that we can extract a wide range of detailed information from news articles, including types of actions, social movement organizations involved, claims made, size, and whether police or protesters used violence. The main disadvantage of this approach is that it is highly labor-intensive and expensive, requiring careful readings of back issues of daily newspapers or a sample thereof. Because of the high cost, researchers must restrict parameters to the number of newspapers coded, particular geographical regions, and specific time periods. This restriction limits the cases to which we can test hypotheses and the quality of the data in terms of comprehensiveness.

The primary goal of this paper and the larger project which it is initiating is to build, test, and validate an automated system for the coding of protest event data from digitalized news sources, using technological advances from computer science and statistics, namely natural language processing. I call this system the **M**achine-learning **P**rotest **E**vent **D**ata **S**ystem, or MPEDS. The aim of MPEDS is to reduce the labor required to generate protest event data and to minimize the biases associated with newspaper coverage of protest events. They will also have reliability rates which are comparable to human coders. The resulting datasets will contain rich information relevant to social movement scholars, include longer-term temporal coverage (including real-time coverage) and introduce the potential of coding for protest events from multiple news sources with worldwide coverage. MPEDS will also be open, available for replication, and extendable by future social movement researchers, and social and computational scientists.

This paper is ordered as follows: I first give a short primer on protest event data and its uses within social movements research. I then introduce the MPEDS, machine learning, and the methodological advances in text as data. I discuss how MPEDS is an improvement over other systems which produce political event data with automated methods. I then outline the components of the MPEDS system – namely the haystack, closed-ended, and open-ended coding tasks. I present evaluation metrics for each part of the system, and in the process compare the suitability of different types of news sources for training the MPEDS classifiers. I then show that many features of MPEDS have comparable reliability to human coders. I close by discussing the future tasks to be accomplished within MPEDS, and suggest implications of the system for social movement research.

2.1 Protest event data

Protest event data is the "who, what, when, where, why, and how" of collective contentious activity. We want to know who is protesting, what claims they are making, who they are targeting, at what time, in what location, and with what methods of protest. Social movement scholars have used protest event data to study a number of significant phenomenon, including the onset of collective ethnic and nationalist violence (Olzak, 1989, 1992; Beissinger, 2002), protest cycles (Tarrow, 1989), the diffusion of ethnic rioting (Spilerman, 1970, 1971, 1976; Myers, 2000), movement responses to police repression (Khawaja, 1994; Earl, Soule and McCarthy, 2003; Earl and Soule, 2006; Davenport, 2010), legislative responses to movement activity (McAdam and Su, 2002), and innovation by social movement organizations (Soule, 2009*b*; Wang and Soule, 2012). Within political science, protest event data (as well as data on other political events) is used primarily for political forecasting of political instability (Goldstone et al., 2010) and the onset of political conflict and violence (Brandt, Freeman and Schrodt, 2011; Schrodt, Yonamine and Bagozzi, 2013). Others have highlighted the rise in attention by political scientists towards civil strife, including civil wars, political violence by non-state actors, as well as protest and political expression (Nardulli, Althaus and Hayes, 2015).

Typically scholars have relied on newspapers as records of political and protest events[1]. Within social movements scholarship, protest event data has usually been extracted from

[1]Official sources such as government and police records are not kept consistently, are contingent on the willingness of a government in sharing their data and how readily accessible those data are, and often don't contain the information which movement scholars are interested. For this reason, only a few datasets have been assembled (Maney and Oliver (2001) for Madison, Wisconsin; McCarthy, McPhail and Smith (1996) for Washington, DC; and Chris Sullivan and Christian Davenport's work on Guatemala).

newspapers for a specified time period in a single or handful of countries, typically from one or a handful of newspapers at most. Tilly, Tilly and Tilly (1975) coded for violent events in France, Germany, and Italy by using national newspapers over the period of nearly a century. Tarrow (1989) coded on collective protest in Italy's main newspaper of record from 1966-1973. Olzak (1992) coded for ethnic collective events in the United States from the *New York Times* from 1877-1914. In their study of "new social movements", Kriesi et al. (1995) coded for protest events from 1975-1990 in four European countries from four newspapers.

Most of these datasets have been collected to support their authors' specific research projects and are thus rarely re-analyzed by other scholars. However, recently scholars have made an effort both to establish a standard methodology for collecting event data and to collect more comprehensive data to be deployed in a variety of movement research. In an effort to establish a common (i.e. not project-specific) method for the collection of event data, Franzosi (2004, 2010) outlines "quantitative narrative analysis," which consists of a formal grammar for documenting historical narratives. Within this grammar, coders must identify the subject, object, and action of an event from historical sources, including newspapers.

Many of these datasets use a handful of news sources, and there is a large body of literature which highlights the differences between news sources nominally covering the same time periods and geographic areas (Franzosi, 1987; Earl et al., 2004). Although there is no perfect measurement of the underlying flow of collective events to provide a basis for comparison, some have suggested that compiling events from many different news sources that vary in location and political slant is the best way to get as close as possible to the true flow of events (Woolley, 2000; Myers and Caniglia, 2004). For example, in his study of collective protest and violence in former Soviet states, Beissinger (2002) uses a wide mix of Western, official Soviet and post-Soviet, and émigré news sources, many of which he obtained from news clipping archives that had been compiled by others[2]. Similarly, Carter

[2] http://www.princeton.edu/~mbeissin/research1.htm#Data

(1983) compiled a comprehensive dataset of urban riots between 1964 and 1971 in the US from multiple sources, including the Congressional Quarterly's Civil Disorder Chronology, the *New York Times*, and the *Washington Post*. The rise of electronic archives of newspapers and the availability of new digital media have made it even easier to access multiple news sources.

2.2 The Machine-Learning Protest Event Data System

The goal of this paper is to introduce and highlight the advantages of my own system, the Machine-Learning Protest Event Data System, or MPEDS. The goal of MPEDS is to provide high quality protest event data using tools from machine learning and natural language processing with little to no human intervention. Before introducing this system, I briefly highlight the growing field of machine learning and data science, and the methods which it introduces. I then review similar automated systems for political event data generation and note how MPEDS improves upon these systems.

Machine Learning and Text as Data

Machine learning can be defined as a set of probabilistic methods that can automatically detect patterns in data and use that information to make predictions in other data (Murphy, 2012). Machine learning methods are often used for classification, ranking, or recommendation. Examples of each include deciding whether Twitter users are liberal or conservative based on their tweets (classification, e.g. Conover et al., 2011), Google's "Priority Inbox" (ranking), and Netflix's suggestions of new products for consumption (recommendation).

Machine learning has become ubiquitous in applications within computer science, and familiarity with its principles and methods is a prerequisite in the burgeoning field of "data science." However, it is only beginning to make inroads within the social sciences, primarily within the field of natural language processing or what has come to be known as "text as data" within political science and digital humanities. The intersection of machine learning and natural language processing has been a fruitful one and has produced a set of common methods and procedures. Within social science, Grimmer and Stewart (2013) provide a good overview of different modes of machine learning, procedures required for treating text as data, and applications within political science. The cultural analysis journal *Poetics* dedicated an issue to topic modeling, a form of unsupervised learning for text, and discusses its implications for social sciences (Mohr and Bogdanov, 2013).

Machine-assisted approaches to political event data have been in use for nearly 30 years, since the inception of the Kansas Event Data System (KEDS; Gerner and Schrodt, 1994) and its progeny (PETRARCH/Phoenix; Schrodt, Beieler and Idris, 2014). More recently, there have been several approaches which incorporate machine learning methods into their pipelines. The SPEED system (Nardulli, Althaus and Hayes, 2015) uses supervised machine learning to help filter out articles which do not contain an event of interest. Croicu and Weidmann (2015) use an ensemble of supervised machine learning classifiers to filter out irrelevant articles in a similar manner to the SPEED project. Neither of these projects, however, attempts to construct a fully automated system. The most significant attempt for a full automated process has been attempted by the Political Conflict in Europe in the Shadow of the the Great Recession (POLCON) project[3]. Wueest, Rothenhäusler and Hutter (2013) and Marakov et al. (2015) have attempted an initial foray into full automated but had limited success. Many of the issues they faced are endemic in this full automation of protest event data extraction, which I will outline further below.

Comparing MPEDS to Other Approaches

MPEDS differs from other automated approaches to producing protest event data in several ways. Like SPEED, it uses a supervised method, meaning humans provide "training" data on which its classifiers are based. And like KEDS, it aims to be open and transparent in its

[3] `http://www.eui.eu/Projects/POLCON/Home.aspx`

data production and pipeline. However, MPEDS differs from other automated event data projects in two major regards: scope of the event and amount of data provided for each event.

Instead of attempting to do many things somewhat well, MPEDS attempts to do one thing very well: identify protest events. In other automated projects, the protest event is ill-defined or subsumed under a more general political event. This has the consequence of both providing a very sparse amount of information for any given protest event (since all political interactions are reduced to a common denominator of information) and by shifting the definition of a protest event such that it fits more neatly into other kinds of political interactions, which has the consequence of forcing the event into a state-centric idea of political interaction. KEDS and its progeny fall victim to the data sparsity problem. Every event is a dyad between two state or non-state actors. Beyond defining actors, targets, and a single political action, no other information is provided about the event. The CAMEO ontology and the SPEED system define protest in a manner that is a poor fit for movements research. The CAMEO ontology used with KEDS is geared towards international relations events – namely, mediation – and not social movement ones (Schrodt, Beieler and Idris, 2014). In addition, CAMEO's event ontology was originally developed to document actions in the Middle East, thus may be skewed in ways that restrict its applicability to other regions. SPEED defines a protest event as an act of "political expression", which includes many of the categories considered as protest by movement scholars, but also includes other behavior, including the publication of dissident media and cultural arts (cartoons, movies, plays).

MPEDS defines a protest event based on an engagement with social movement theory and a survey of hand-coded datasets within the social movement literature. It also differentiates itself from other automated projects by attempting to find a good medium between the sparse dyadic data of KEDS and the hand-coded and textured data produced by a hand-coded project. MPEDS provides a number of variables on protest events which have been of historical importance to movement scholars. The system is structured in this way such that a more fully automated solution can process news sources with minimal human intervention. Lastly, the MPEDS system, the human coder web interface, and the data produced by MPEDS will be offered as open-source and distributed publicly. In addition, events will include an audit trail, such as a URL (if available), and the article title and news source, such users can identify the source text of the article and reassess the data on a qualitative basis. MPEDS is thus oriented to produce data which is primarily of interest to movement scholars, both by definition of the event and by the information which is included in each record.

MPEDS Architecture

Within MPEDS, there are three discrete tasks: the haystack task, the closed-ended coding task, and the open-ended coding task. The haystack task discerns whether a document mentions a protest event. I call this the haystack task because the problem is largely imbalanced – articles that mention protest are rare relative to the total number of articles in any given news source. The closed-ended task attempts to classify several variables which can take on a discrete number of values. I focus on three variables: the dominant *form* of the protest, the main *target* of the protest, and the main *issue* of the protesters. The task is, for each of the documents identified as mentioning a protest event, is to classify the document for each of these variables. The final task is to pull out relevant information of the protest, the "open-ended coding" task. I focus on a protest's size, its location, and the name(s) of social movement organizations involved (if any).

The first two tasks can be treated as multiclass document classification problem; the last task can be treated as a named entity recognition and pattern matching task. Table 2.1 summarizes the tasks and the methods of data generation. These tasks will be outlined in detail below.

The MPEDS project is also collecting news text and coding data by hand in order for the

Variable	Task	Method
Contains protest?	Haystack	Binary classification
Issue	Closed-ended coding	Multiclass classification
Form	Closed-ended coding	Multiclass classification
Target	Closed-ended coding	Multiclass classification
Size	Open-ended coding	Pattern matching
Location	Open-ended coding	NER + Dictionaries (gazetter)
Organization names	Open-ended coding	NER + Dictionaries

Table 2.1: Variables and methods of classification

system to use as a training data. Using a web interface, coders must first discern whether an article contains a protest event (the haystack task) and then highlight the text in which variables of interest are present. Although many of the variables (e.g. claims) are not explicit in the text, we must rely on the text itself to produce variables of interest. After this "first pass" of coding, articles which are candidates for event coding are passed to a "second pass", in which coders disentangle multiple events in a single article, categorize forms, claims, and targets into discrete categories, and double-check the coding for specific locations, dates, social movement organizations, and crowd sizes.

The main aim in creating this hand-coded dataset is not comprehensiveness of coverage over a particular time period or particular news source. The goal is incorporating enough protest articles from a diverse number of sources in order to account for all the different ways in which a news source may talk about protest activity. Different sources possibly use different words and word combinations to talk about a protest event. Therefore, we code for news sources which may have stylistic differences in reporting, rather than simply spatial variation. Figure 2.1 illustrates the entire MPEDS pipeline, including the process of incorporating new training data. The methodology is as follows:

1. Select a number of news sources of interest. Include variation based on location, audience (national, regional, international), format (newspaper, news wire), and time period in order to account for any period effects in language.

2. Sample an adequate amount of articles to generate sufficient protest articles for building training and test sets. Given prior testing and other machine learning projects (e.g. Hopkins and King, 2010), this is between 50 to 100.

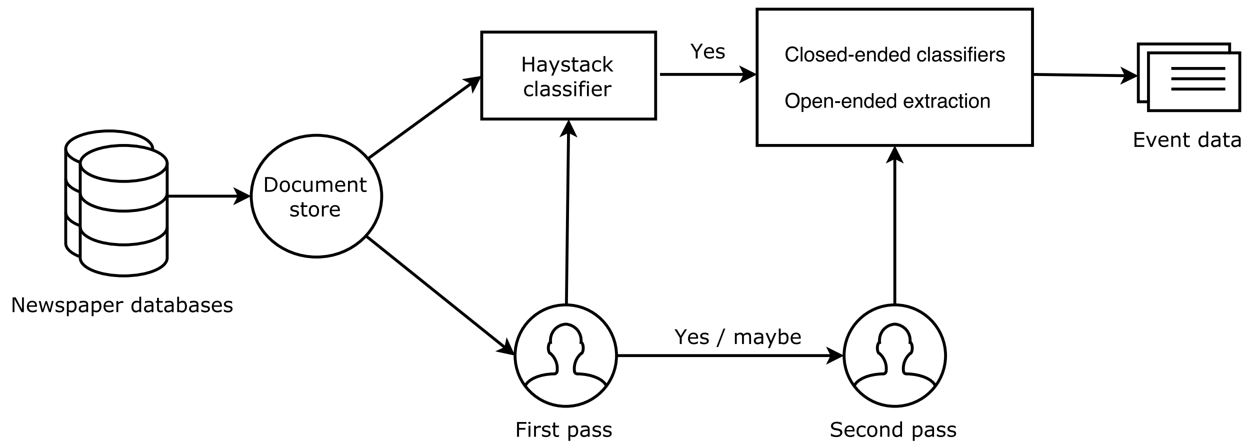


Figure 2.1: MPEDS pipeline with training.

3. Search a news database (e.g. Lexis-Nexis) with a broad search term that includes all uses of protest-related words. This helps to reduce extraneous articles.

4. Pipe these articles to a first pass of coding in which coders decide whether an article involves a protest or not, and highlight relevant parts of the text. This task filters out over 80% of articles.

5. Send all articles which are labeled as protest to a second pass in which coders construct discrete events from articles.

Potential Training Data: Dynamics of Collective Action

MPEDS originally attempted to use an existing protest dataset as training data. The Dynamics of Collective Action (DoCA) dataset is, to date, the largest protest event dataset of events occurring in the United States[4]. To generate this dataset, humans hand coded articles from the *New York Times* from 1960 to 1995. This resulted in a dataset of nearly 21 thousand unique events. DoCA includes any event which meet the following criteria: (1) collective acts; (2) public actions; (3) protest actions (e.g. not a fundraising event or a closed group meeting); and (4) are making a specific claim or grievance about the desirability to change society[5]. In addition to coding for protest, DoCA includes ethnic/racial conflict

[4]The dataset and accompanying codebooks can be found at `http://www.stanford.edu/group/collectiveaction/cgi-bin/drupal/node/1`.

[5]`http://www.stanford.edu/group/collectiveaction/BRIEF%20EVENT%20GUIDE.docx`

events and lawsuits related to social movement activity. However, in order to have a more strict specific definition of a protest event, we excluded these events from our analysis. Each event is coded for a comprehensive list of variables, including date, size, location, a qualitative description of participants and events, claims, forms of protest, protest targets, initiating groups involved, presence of violence, and presence of police.

I treat DoCA as a potential training set for the haystack task. DoCA seems well-suited for this purpose, given the large number of events in the dataset and the number of events which can be matched to their original articles in the *New York Times*. To match events to source articles, I use the New York Times Annotated Corpus[6] obtained by the University of Pennsylvania Linguistic Data Consortium (LDC-NYT), a machine-readable dataset of 1.8 million *New York Times* articles from 1987 to 2007. DoCA contains a total of 3,570 contentious collective action events during this period that should have a corresponding article in LDC-NYT, that is, from 1987 to 1995. In practice, however, I have found that not all records in DoCA could be matched to a source article in LDC-NYT, either due to a malformed transcription into LDC-NYT, DoCA coders sourcing the event from an AP wire report that does not appear in LDC-NYT, or some updated or otherwise changed title in LDC-NYT. However, with minor data cleaning I matched about 88% (3,214 of 3,570) of the articles in DoCA to their source texts.

From 1987 to 1995, there were nearly 820 thousand articles in the *New York Times*, while there are only about 400-500 events per year within DoCA. Following Leetaru and Schrodt (2013), I filtered out a handful of common titles related to business, finance, and sports (e.g. "Business Report") which are not be relevant to the project. The LDC-NYT also contains a field which lists a taxonomical classification when indexed online, and of this I exclude Business, Finance, Sports, and Classified categories. I also exclude Weddings and Book Reviews based on the *New York Times* index field. This filters out more than half of the articles which we can assume do not mention protests. For the final filter, I sampled all

[6] `https://catalog.ldc.upenn.edu/LDC2008T19`

articles from the LDC-NYT on each date in which there was a record in DoCA using a broad search string, described in Appendix A.3. In our final count, we have just over 50,000 potential protest-related articles.

Data generated from the MPEDS project

The MPEDS project has collected news text data from over a dozen sources, including several local and national US newspapers, and news wire services. I focus on several sources across all geographical coverage areas. Each source is displayed in table 2.2, along with the number of articles used in its training set, the number of articles found to contain a protest by human coders, and that value as a percentage of total articles. All sources (except for DoCA, which runs from 1987-1995, and NYT, which is from 1996-2007) were sampled from the beginning of 1995 to the end of 2010. We used the search string specified in Appendix A.3 to filter out articles. From each source, we drew a sample of 150 dates which was stratified to oversample on Sunday editions. National news sources include the *New York Times*, the *Washington Post*, and *USA TODAY*, local sources include the *Austin American-Statesman*, *Omaha World Herald*, and the *Atlanta Journal-Constitution*, and news wires include Agence France-Presse and the Associated Press. The *New York Times* data were drawn from the LDC-NYT dataset and news wires from the Annotated English Gigaword dataset[7], also provided by the Linguistic Data Consortium. The other sources were downloaded from Lexis-Nexis. Following Nardulli, Althaus and Hayes (2015), I stored all articles and related metadata in an Apache Solr document store for quick access, version control, and indexing.

MPEDS defines a protest event in the following manner: the event must involve some form of claims-making and grievance expression, have sufficient information for coding, i.e. location and date, occur in public, and include at least some non-institutional actor. A full definition of acceptable (and non-acceptable) protest events is located in Appendix A.2. Coders went through at least one month of training which included weekly team

[7] `https://catalog.ldc.upenn.edu/LDC2012T21`

Source	Total	Protest	Protest %
Agence France-Presse (AFP)	4131	678	16.41
Associated Press Worldstream (APW)	3891	483	12.41
The Washington Post (WPO)	3413	332	9.73
USA TODAY (USA)	903	84	9.3
Austin American-Statesman (AUS)	663	38	5.73
The New York Times (NYT)	3051	162	5.31
Omaha World Herald (OMA)	784	32	4.08
The Atlanta Journal and Constitution (ATL)	1350	55	4.07
Dynamics of Collective Action (DoCA)	50266	1079	2.15

Table 2.2: Descriptive statistics on news sources for training datasets. Name abbreviations are in parentheses.

discussions and reviews of reliability reports generated from the project data. I included as a protest any article in which over 50% of coders labeled as such.

Table 2.2 reports the number of protest articles in each data source as a percentage of the total sample. DoCA has the lowest number of 2.15% and NYT has fourth lowest at 5.31%. Theoretically, these two values should be the same. This either indicates underreporting by DoCA coders, overreporting by MPEDS coders, or a significant change in the rate of reporting protest events between the 1987-1995 and 1996-2007 period. Each local source (ATL, OMA, AUS) has a protest article occurrence of less than 6%. WPO and USA, the other national newspapers, have rates of protest articles of 9.73% and 9.3%, respectively. The news wire services report the highest percentage of protest articles, 12.41% for the APW and 16.41% for the AFP.

2.3 Haystack coding

The haystack task itself proved to be one of the most difficult parts to train, tune, and to validate of the whole of MPEDS. It's of little surprise that many of the attempts to automate the creation of protest event data have stopped after carefully tuning a set of classifiers or dictionary rules which are able to adequate capture a good deal of the events of interest. This is due to the fact that the social object of the "protest" is itself heterogeneous, difficult to define, and requires explicit boundaries to separate it from routine crime, sport hooliganism, terrorism, or other forms of political violence. Indeed, Hutter (2014*a*) notes how the definitions of protests seem to shift with the focus of the researcher or the specific project. The MPEDS project has sought to be question-agnostic in our own definition, but this naturally does not prevent any of our own intellectual and personal preoccupations from slipping into the analysis. In this section, I outline the steps taken towards developing the haystack task, including text preprocessing, selecting sources, and evaluation.

Preprocessing and evaluation

Article texts went through a series of pre-processing procedures before being used in the machine learning system. They were converted to lowercase and stripped of punctuation and stop words (e.g. common connecting words like "the", "a"). I converted words in the article to numerical representation (a series of feature vectors in machine learning terms) using the term frequency-inverse document frequency metric, or tf-idf. This metric is a measure of word prevalence for word i (w_i); it is calculated by number of times w_i appears in a document divided by the number of times w_i appears in the whole corpus of documents.

I evaluate the accuracy of the system by using metrics of precision and recall from the machine learning literature. These metrics are based on the number of true positives (TP), or correctly classified documents, compared to those which are false positives (FP) or false negatives (FN). Precision can be defined as the fraction of documents correctly classified from the set of all the documents in the class of interest (Equation 2.1), while recall can be defined to the fraction of documents correctly classified from the set of all documents (Equation 2.2). Maximum precision would indicate the absence of false positives, while maximum recall would indicate the absence of false negatives. Precision and recall are thus analogous to the Type I (incorrect rejection of a true null hypothesis) and Type II errors (failure to reject a false null hypothesis), respectively. Precision and recall are tradeoffs by definition.

$$precision = \frac{TP}{(TP + FP)} \tag{2.1}$$

$$recall = \frac{TP}{(TP + FN)} \tag{2.2}$$

I used F_β-scores (or F-score, Equation 2.3) to evaluate the overall model. This score is the harmonic mean of the recall and precision. In the haystack task, I use the F_2-score, which weights the recall with more importance. Otherwise, I use the F_1 score, which weights them equally.

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \tag{2.3}$$

Within the machine learning literature, there are no hard numerical cutoffs on the acceptability of any one of these metrics. The cutoff is more or less application-dependent. If a researcher is more interested in retrieving documents of interest with some level of noise, prioritizing recall should be more important. Conversely, if the researcher wants to identify the most relevant documents and risk losing some in the process, precision should be prioritized. For this paper, I use 0.65 as a lower boundary for an acceptable F_β-score.

Classifiers were tested using k-fold cross-validation with $k = 3$. K-fold cross-validation withholds a single slice or "fold" of the data for testing while training on the other $k - 1$ folds. For the haystack classification, I used an ensemble classifier, which has been used with success in other political and event analysis (e.g. Grimmer, Messing and Westwood, 2014; Croicu and Weidmann, 2015). Ensemble methods work by applying several different classifiers to the same dataset and giving each classifier a "vote" on the article's classification. After testing several combinations, I obtained the best results combining a support vector machine (SVM) classifier with a linear kernel, a logistic regression (LR) classifier, and three

Source	All-P	Own-P	All-R	Own-R	All-F2	Own-F2
afp	0.56	0.67	0.84	0.63	0.76	0.64
apw	0.51	0.74	0.88	0.68	0.77	0.69
atl	0.58	0.35	0.54	0.54	0.55	0.49
aus	0.67	0.41	0.73	0.57	0.72	0.52
doca	-	0.33	-	0.58	-	0.50
nyt	0.58	0.54	0.61	0.53	0.61	0.53
oma	0.60	0.20	0.47	0.34	0.49	0.29
usa	0.45	0.35	0.55	0.49	0.52	0.45
wpo	0.57	0.65	0.60	0.42	0.59	0.45

Table 2.3: F_2 score per test source and each training source. Own-* is the metric using only the same source in training. All-* is the metric using all sources in training.

stochastic gradient descent (SGD) classifiers with different loss functions: the hinge loss function, the perceptron loss function, and the Huber loss function. For features, I used the tf-idf metric on unigrams and bigrams, that is, one- and two-word combinations. I discuss alternative model specifications and feature selection below.

In practice, multiple sources would be used train the haystack classifier, rather than just one. This is because we want to be able to capture events in a variety of sources, not just one or two major ones. To this end, I first assess classifier performance based on a training set composed of the same source as the test set. I then move to evaluation of classifiers based on every combination of two sources, and conclude with classifiers based on all sources.

Results

Results from the haystack task are reported in Table 2.3, which reports the precision, recall, and F_2 scores for the classifiers using its own source and all sources. Since there is only one combination of training choices for both the classifier based on all sources or its own source, only those two are reported. Figure 2.2 plots the distribution of F_2 scores by classifiers trained on own source, pairs of sources, and all sources. For DoCA, I only evaluate the classifier based on its own articles. I do this to illustrate the adequacy of using DoCA itself as a training set.

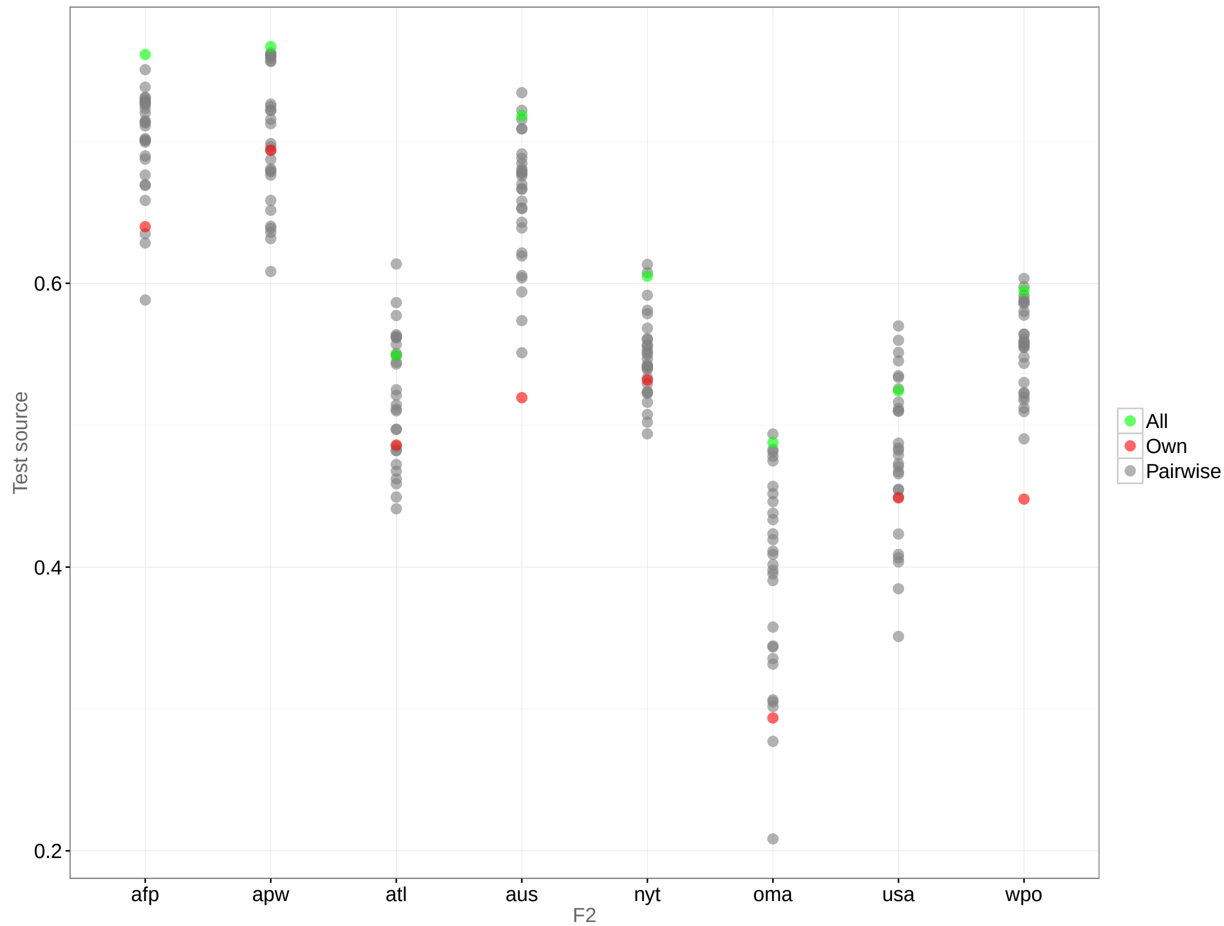


Figure 2.2: F_2 scores for classifiers trained on all, pairs, and their own sources

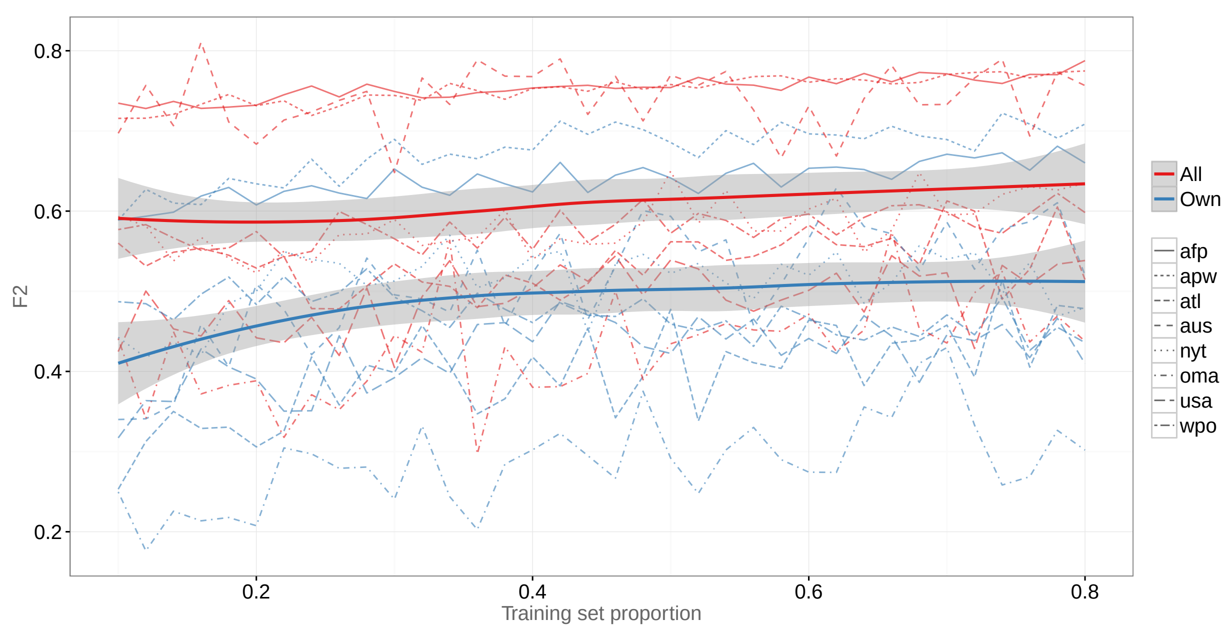


Figure 2.3: F_2 scores for own and all sources, by training proportion size

The most notable things from Table 2.3 seems to be, first, the large disparity between different sources, and second, the large gains when using the classifier trained on all sources. Results for the own classifier range from 0.29 to 0.69. The classifiers which perform the worst are the local Omaha paper (0.29), then two national sources – the Washington Post (0.45) and USA TODAY (0.45). While the low accuracy for the Omaha paper could be attributed to the small amount of training data used in training, the larger national papers do not have that issue. The best performing own classifiers are the news wire services, AFP (0.64) and APW (0.69). In the middle is the DoCA (0.50) and the New York Times (0.53). Their F-scores are very similar, but there is a large gap between their precision (DoCA's 0.33 compared to NYT's 0.54) and recall (DoCA's 0.58 compared to NYT's 0.53). This would seem to indicate that there is a large number of false positives which are being reported from DoCA. These results merit a small note.

In a separate analysis, I sampled 47 articles which had been marked as false positives by the classifier trained on DoCA data. Of these 50 articles, 31 of them were "false false positives", that is, they were shown to be articles that should have been in DoCA by the project's own criteria but were not. This result highlights a rather large margin of error in DoCA, introduced either by coder fatigue or technological error.

The incorporation of more training sources seems to universally provide for better accuracy. In Figure 2.2, the gray points represent F-scores for classifiers trained on pairs of sources. In a few cases, these classifiers decrease accuracy, especially with the NYT, APW, and USA sources. This typically is the case when its own source is not part of the pairwise source. But on the whole, the pairwise comparisons provide a net positive.

In nearly all cases, incorporating all sources provides for the best or one of the best classifiers, noted by the green points. All classifiers see an increase in F_2 score. The largest gains are seen by two local papers - AUS and OMA. Each source sees at least an increase of 0.06 in F-score. The floor for accuracy is now 0.49 (OMA) and the maximum is 0.77 (APW). The addition of more sources therefore provides more heterogeneity in reporting and increases classifier power by a large factor.

It's worth nothing that the news wire services have the best accuracy of any of the news sources. News wires have the highest proportion of protest articles in the dataset. If one is interested in capturing the most events on a worldwide basis – rather than detecting events which are "socially significant" - then it seems like news wires would be a good place to search. Indeed, other event data researchers have noted the virtues of news wire services as well, despite their other drawbacks (Schrodt, Davis and Weddle, 1994; Schrodt, Simpson and Gerner, 2001).

Figure 2.3 reports the increase in F_2 score as more of the training set is used for training. The solid line is a LOESS regression across all news sources tested. There is a consistent pattern of most sources which use all training sources for classifier as having better accuracy. On average, even a training set using 20% of data for training still does better on the whole than using 80% of data for training for the individually-sourced classifiers.

While the haystack task seems straightforward from the outset, since it is a simple matter of detecting whether an article mentions a protest or not, the task is more complicated than it seems at first glance. The variation in accuracy with a similar classifier across multiple news sources seems to point to some kind of fundamental aspect of the news text which impedes a simple binary classifier. Using multiple news sources with an ensemble of classifiers seems to be the best strategy.

In terms of feature selection, I chose to use a simple bag-of-words approach. This approach is computationally inexpensive and can be accomplished with many well-defined tools. However, it may be possible to use part-of-speech tags of words to discern between different uses of words. For instance, there are the semantic differences between $\mathit{March}_{\mathrm{NNP}}$ (the month), $\mathit{march}_{\mathrm{NN}}$ (the noun and actively moving group of people), and $\mathtt{march}_{\mathrm{VBZ}}$ (the 3rd person singular present verb and common protest activity)[8]. There have also been other

[8]Subscripts are adopted from the Penn Treebank's part-of-speech tags: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`.

great advances in word sense disambiguation with the release of the word2vec tool[9], which is very good at finding similar words based on word context and constructing analogies between sets of words. Both part-of-speed tagging and word2vec are more computationally intensive in the preprocessing stage, however.

One solution I attempted was reducing the dimensionality of the feature space using Latent Dirichlet Allocation (Blei, Ng and Jordan, 2003). Latent Dirichlet Allocation (LDA or topic modeling) is a hierarchical Bayesian model which allows documents to belong to multiple classes (or topics). Each word in the document has a distribution over the topics and is thus more flexible than supervised classifiers. However, this dimensionality reduction approach did not yield better results for the haystack task. There may be ways to successfully apply other unsupervised methods to the haystack classification task, but with the combination of several sources the results are sufficient for our purposes.

2.4 Closed-ended coding

For the closed-ended coding task, I created classifiers for each of the variables which take discrete values: target, issues, and forms. As noted by the Figure 2.1, the training values were sourced by second-pass coders, and most articles were coded at least twice in the process. Coders could assign more than one value to an article and therefore compound variables constitute their own class. For instance, rallies/demonstrations and marches tend to have a high rate of co-occurrence, so one frequently used compound value for the form variable is "Rally/demonstration-March". I limited the cardinality of these compounds to two, given that there is a very long tail on possible combinations. A value was also not used in the cross-validation set if it did not appear at least than 30 times.

In order to decide which values to use, I constructed a set of coding rules for inclusion of training data values.

[9] https://code.google.com/p/word2vec/

	P	R	F1	N
0: Blockade/slowdown/disruption	0.58	0.07	0.11	46
1: Boycott	0.87	0.35	0.50	54
2: Hunger Strike	0.95	0.60	0.73	53
3: March	0.64	0.38	0.47	191
4: Occupation/sit-in	0.50	0.06	0.10	35
5: Rally/demonstration	0.66	0.93	0.77	809
6: Rally/demonstration-March	0.33	0.11	0.16	72
7: Riot	0.63	0.34	0.43	82
8: Strike/walkout/lockout	0.84	0.85	0.85	273
9: Symbolic display/symbolic action	0.59	0.17	0.26	52
10: _none_	0.00	0.00	0.00	13

Table 2.4: Form accuracy metrics

1. *Total agreement*: If there was total agreement, i.e. every value is the same for all coders, then use all values as expected.
2. *Partial agreement*: This is if coders agreed on one or more values but not others. Two cases apply here: if there are more than two coders and there are values which have taken on a majority vote, use the majority vote (e.g. coder 1: march, coder 2: march, coder 3: rally, use march). Otherwise, use the intersection of all coders.
3. *None vs. any*: One coder hasn't coded a value or has coded None of the above, while the other coder has. Use the non-none value.
4. *Total disagreement*: In the last case, coders do not agree on anything. Discard the case and do not use it in the analysis.

After testing several different classifiers for each variable, I settled on different classifiers for each variable. For form, I used the LR classifier, for issue, an SGD classifier, and for target, an ensemble voting classifier based on SVM, SGD, and LR. Each classifier used a "One vs. the Rest" approach, in which a separate classifier was trained for each value such that the classification task assessed fit for that particular value versus all other values.

Like in the haystack task, I use a 3-fold cross-validation method to assess the classifier accuracy. Tables 2.4, 2.5, and 2.6 report the precision, recall, F_1, and number of cases across all folds for each of the closed-ended variables. I chose to use the F_1-score for this task because there is no theoretical reason in which recall should be more important to precision in this task. Tables 2.7, 2.8, and 2.9 report the confusion matrices for each of the

	P	R	F1	N
0: Abortion	0.94	1.00	0.97	30
1: Anti-colonial/political independence	0.48	0.52	0.50	67
2: Anti-war/peace	0.61	0.70	0.65	132
3: Civic violence	0.54	0.21	0.30	43
4: Criminal justice system	0.64	0.57	0.60	96
5: Democratization	0.65	0.73	0.68	172
6: Economy/inequality	0.61	0.55	0.58	130
7: Environmental	0.71	0.91	0.79	64
8: Foreign policy	0.53	0.37	0.42	82
9: Human and civil rights	0.52	0.27	0.35	78
10: Immigration	0.81	0.81	0.81	37
11: Labor & work	0.80	0.92	0.85	341
12: Political corruption/malfeasance	0.52	0.48	0.49	94
13: Racial/ethnic rights	0.50	0.60	0.53	81
14: Religion	0.82	0.86	0.84	115
15: Social services & welfare	0.49	0.39	0.41	52
16: _none_	0.31	0.16	0.21	85

Table 2.5: Issue accuracy metrics

	P	R	F1	N
0: Domestic government	0.80	0.92	0.86	1097
1: Foreign government	0.78	0.74	0.76	348
2: Individual	0.53	0.09	0.15	56
3: Intergovernmental organization	0.81	0.60	0.68	99
4: Private/business	0.80	0.73	0.76	233
5: University/school	0.42	0.19	0.26	37
6: _none_	0.00	0.00	0.00	24

Table 2.6: Target metrics

	0′	1′	2′	3′	4′	5′	6′	7′	8′	9′	10′
0	3	0	0	1	0	35	0	1	5	1	0
1	0	19	0	1	0	29	0	1	4	0	0
2	0	0	32	0	0	17	0	0	3	1	0
3	0	0	0	72	0	98	11	2	7	1	0
4	0	0	0	2	2	28	0	0	2	1	0
5	3	1	1	16	0	752	6	11	17	2	0
6	0	0	0	16	0	46	8	0	1	1	0
7	1	1	0	1	0	51	0	28	0	0	0
8	0	0	0	3	0	37	0	0	233	0	0
9	0	0	1	1	1	35	0	1	4	9	0
10	0	1	0	0	0	12	0	0	0	0	0

Table 2.7: Form confusion matrix

	0′	1′	2′	3′	4′	5′	6′	7′	8′	9′	10′	11′	12′	13′	14′	15′	16′
0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	35	4	0	1	2	4	0	7	3	0	0	1	4	4	1	1
2	1	5	92	1	5	2	3	3	3	1	0	1	3	4	5	2	1
3	0	0	7	9	0	3	3	0	1	3	0	3	4	4	2	1	3
4	0	2	6	2	55	6	2	2	3	3	0	1	1	3	2	3	5
5	0	3	4	0	3	125	1	3	1	0	2	4	17	3	2	1	3
6	0	4	6	0	1	4	72	3	0	1	0	29	4	1	0	5	0
7	0	0	2	0	1	1	1	58	0	0	0	0	0	0	0	0	1
8	0	12	7	0	3	4	3	3	30	3	0	5	0	5	4	0	3
9	0	1	4	1	9	7	4	1	3	21	4	4	4	6	1	4	4
10	0	1	0	0	0	1	0	0	1	2	30	2	0	0	0	0	0
11	0	0	2	0	0	2	14	1	1	0	0	314	2	3	0	2	0
12	0	0	4	0	1	25	0	0	1	3	0	5	45	3	2	0	5
13	0	1	2	1	2	3	4	2	6	2	0	3	1	49	1	0	4
14	0	4	3	1	1	1	0	1	1	0	0	1	0	3	99	0	0
15	0	0	1	0	2	2	3	1	1	0	0	13	2	2	0	20	5
16	1	5	8	2	2	6	4	5	4	1	1	10	4	13	0	5	14

Table 2.8: Issue confusion matrix

	0′	1′	2′	3′	4′	5′	6′
0	1005	47	2	3	36	4	0
1	77	259	0	9	3	0	0
2	46	3	5	0	0	2	0
3	22	16	0	59	1	1	0
4	54	4	2	2	170	1	0
5	26	0	1	0	3	7	0
6	19	3	0	0	1	1	0

Table 2.9: Target confusion matrix

variables. A confusion matrix displays the predicted class of the document compared to its actual class. The column names are marked x' to denote the cases which were predicted as class x. The rows are the actual class. The value on the diagonal is the number of documents which were classified correctly. So for instance, in table 2.7, the value in row 5 (rally/demonstration) and column 7′ (riot) is 11, which means 11 articles human coders labeled as rally/demonstration were coded as riots by the classifier. I will discuss each of the closed-ended variables in turn.

For form, F_1 ranges from 0.11 to 0.85 for all non-none categories. Only three classes have an F_1 over 0.5: hunger strike, rally/demonstration and strike/walkout/lockout. It is not for want of training data either, since march, with 191 cases, has an F-score of 0.47. In the confusion matrix, the most noticeable thing is that misclassification of events towards 5′, rally/demonstration. The form classifier does not do well to distinguish between the rally and other types of events. All-in-all, the classifier mislabels 342 events as rallies (not including the compound category, rally/demonstration-march). On the other hand, few events are mislabeled as a strike/walkout/lockout, the second-most populous category.

What explains this classification error? It may be the case that since rally is the overwhelming form of protest event, everything is very much tinged with the same kind of language. Frequently, when there is an occupation, a boycott, or a march, it is accompanied by a rally. This result is verified by DoCA, where rallies occur in at least 21% of events[10].

[10]http://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/node/17

The intention behind creating compound variables is to capture the co-occurrence of forms. But even with that, it doesn't seem like an automated method is able to distinguish between more nuanced types of contentious action, save the hunger strike and the labor strike. But as I will note below, these rates of error seem similar to that of human coders.

For issue, F_1 ranges from 0.30 to 0.97 for all non-none categories. There are several categories with an F-score above 0.8: abortion, immigration, labor & work, and religion. Notably, abortion is retrieved with perfect recall and near perfect precision with the minimum number of cases allowed for inclusion (30). Several more classes have F-scores equal or above 0.6: anti-war/peace, criminal justice system, democratization, and environmental. Below that are three are in the 0.5 decile: anti-colonial/political independence, economy/inequality, and racial/ethnic rights. The last categories are below 0.5: civic violence, foreign policy, human and civil rights, political corruption/malfeasance, and social services & welfare.

On the whole, errors aren't as biased towards any one category as they are in the case of protest forms. In the confusion matrix, no single category seems to be driving misclassification. Select pairwise misclassifications seem to be driving the error. Articles labeled economy/inequality are most frequently misclassified as labor & work (29), which makes substantive sense. Articles labeled as democratization are most frequently misclassified as political corruption/malfeasance (17) and vice versa (25). This seems to follow from the logic that many democratization movements are often driven or at least in response to political corruption by regime elites. But otherwise, there is no single category towards which the classifier exhibits a systematic misclassification.

Lastly, for target, F_1 ranges from 0.15 to 0.86 for non-none variables. Domestic government, foreign government, intergovernmental organization, and private/business all have F-scores above 0.65. Below that, university/school has a poor F-score of 0.26 and individual has a very poor F-score at 0.15. In the confusion matrix, we see the same systematic misclassification towards domestic government for all categories: 244 articles are

	P	R	F1
Form	0.68	0.69	0.64
Issue	0.64	0.65	0.63
Target	0.77	0.79	0.77

Table 2.10: Weighted accuracy metrics for all closed-ended variables.

misclassified as such. Like rally/demonstration, this represents a bias towards the targets of protest overall. DoCA again validates this result: more than 51% of events in DoCA are targeted towards the domestic state.

Table 2.10 displays the weighted average of accuracy metrics for all closed-ended variables, weighted by the number of cases within each label. Target has the highest F_1 at 0.77, driven mostly by the very large number of cases which are correctly classified as domestic government. Form has an F_1 of 0.64, mostly driven by the prevalence of the rally/demonstration. Issues has an F-score which is marginally worse (0.63), but as noted in the tables above, the classifier does a reasonably well job given the number and heterogeneity of types of issues.

On the whole, these results are promising. While not perfect, the classification performs reasonably well for the task as hand. One outstanding question is whether these classifiers would perform as well as human coders at the same task. One virtue of MPEDS which makes it preferable to that of human coders is how we can code articles at scale. If MPEDS's mistakes are similar to those of human coders, then we could still use it with similar confidence and with significantly more speed.

In the second-pass coding process, most if not all articles were coded by more than two coders. Because of this, we can compare the systematic disagreements of human coders with the disagreements between MPEDS and filtered ground-truth values. This may indicate that the machine learning algorithm is doing about as well as human coders, a claim (King and Lowe, 2003) note in their automated event data system.

As a rough test of this, for all articles which had two or more coders, I generated an

	r	p
Form	0.513	0.158
Issue	0.794	0.000
Target	0.896	0.006

Table 2.11: Pearson correlation and p-values for coder agreement index and MPEDS F_1.

index which captures coder agreement. For total agreement between coders, the article receives a score of 1 for that variable. If coders have partial agreement, then the article receives a score of 0.5. If there is total disagreement, the article receives a score of 0. I summed scores and normalized them by the number of times each value was selected by a human coder. Like in the training data, I excluded those which values which do not appear at least 30 times. I then generated Pearson correlation coefficients between this coder agreement index and the F_1-score of the variables, as well as their corresponding p-values. These values are reported in Table 2.11.

The claim that these classifiers are doing as well as human coders would seem to hold for the issue and target variables. Issue has a correlation of 0.794 with $p < 0.001$ and target has a correlation of 0.896 with a $p < 0.01$. For form, however, the correlation is 0.513, but the p-value does not meet the 0.05 criteria for statistical significance. This provides us with more evidence that we can use these classifiers with confidence as a replacement for human coders.

2.5 Open-ended coding

While the haystack and closed-ended coding tasks classify the article from sets of known categories, the process of extracting text from the document which relate to a variable of interest involves identification of relevant entities and parsing correct information from candidate text snippets.

MPEDS focuses on three open-ended variables of interest: size, social movement organization, and location. Size is important as a metric of protest intensity and has been noted to be a predictor of news bias (Oliver and Maney, 2000). More recently, Biggs (2016) has noted that getting protest intensity correct in protest event datasets is more theoretically important than getting frequency counts right. Thus size is an important variable to code and code correctly. Social movement organizations are significant insofar as the large and still growing literature on the importance of movement organizations, initially motivated by the resource mobilization perspective (McCarthy and Zald, 1977; Zald and McCarthy, 1979) and continuing throughout the current focuses in movement scholarship. Since the establishment in resource mobilization, the focus on social movement organizations has gained insights from organization theory (Davis et al., 2005) and become a standard category for protest event analysis. Obtaining location is simply necessary for us to locate a protest event in space. If there is a date, form, and issue for an event, it does not do us well unless we know where it is.

Although each of these can be extracted from the same text, automating the extraction of each variable has its own set of challenges. There is a significant subset of machine learning research which focuses on labeling sequences (known as sequence classification), such as genomic data and sequences of strings (Xing, Pei and Keogh, 2010). Within natural language processing, the process of identifying entities in text is known as named-entity recognition (NER). Several popular tools have been made to accomplish NER, including Stanford CoreNLP (Finkel, Grenager and Manning, 2005) and MALLET (Sutton and McCallum, 2006). Most modern NER systems use some sort of probabilistic model for text classification. For both Stanford CoreNLP and MALLET, those are conditional random fields (CRF) and hidden Markov models (HMM). Although theoretically these tools could classify any annotated text for a snip of specific text, the most common NER classifiers have been trained to find three entities: organization, person, and location. Luckily, that is two of the three variables we're interested in. Unluckily, however, additional work is needed to distinguish between a social movement organization and any other kind of organization, including organizations the protest may be opposing. Similarly, we need to be able to distinguish between locations in which the protest has occurred and other locations which appear in the article.

Size presents another host of issues. Although numbers and numerical words are relatively easy to find within texts, separating out numbers which represent protest size estimates from other numbers is more difficult. In a typically news article on a protest event, numbers may include not only crowd size, but also numbers arrested, injured, and killed, and crowd size estimates from a protest occurring at an earlier point in time. To complicate matters, estimating crowd size itself is not a very exact science and is contested depending on the social position of the person reporting the number relative to the protest (McPhail and McCarthy, 2004). This contestation often plays out within the news article itself with writers often reporting both organizer and police estimates. Each of these variables requires a particular set of rules for extraction. Below I describe methods of size, organization, and location extraction.

Size. The pseudocode in Figure 2.4 outlines the steps behind detecting protest event sizes. In words: the algorithm anchors the size in a set of digits or number word, then infers if it is a protest size by the context around it. There is often a word which denotes a subject after the numerical term. Some of these are dead giveaways (protester, activist) but most of the time they are other, more general words which refer to a kind of individual (e.g. people, mothers, Egyptians). I created a regular expression to capture a wide swath of these terms and also created a comprehensive list of ethnicities and nationalities from Wikipedia[11]. Before the numerical word may be another number which is a descriptor (e.g. hundreds of thousands). If no subject word had been identified after the word in question, a word occurring before the number may denote a size (e.g. crowd of 10,000). I then return the all the sizes found in the article. Sizes are then resolved to a numerical estimation of the prose description. This estimation is intended to represent scale rather than exact amounts.

[11]"List of contemporary ethnic groups" `https://en.wikipedia.org/wiki/List_of_contemporary_ethnic_groups` and "List of people by nationality" `https://en.wikipedia.org/wiki/Lists_of_people_by_nationality`

```
tokens <- split text by spaces
sizes  <- empty list
for t in tokens
  size <- None
  if t is in digits or t is a number word
    ## this looks for following words
    for f in t + 5
      if there is another number within three tokens; skip
      else if there is not a restricted verb
        if t is in protest subjects
          size <- t
        else if t is a subject word and there is a protest verb in tokens after t
          size <- t
    ## this looks for preceding words
    for p in t - 4
      if size is not set
        if there is a group word ## e.g. crowd of, group of
          size <- t
        else if there is a protest verb in tokens after t
          size <- t
      else
        if there is another number ## this is a descriptor, e.g. four hundred
          size <- size + t
  append size to sizes
return sizes
```

Figure 2.4: Size extraction algorithm in pseudocode.

For instance, “tens of thousands” turns into 10,000, “dozens” turns into 10, and “several” turns into 2.

The previous metrics of precision, recall, and F-score are not appropriate because of the resolution into a continuous variable, thus size accuracy is evaluated both by finding exact matches and calculating the mean Jaccard index across articles. The Jaccard index is calculated by taking the intersection of the set of values identified by a human coder and the set of values identified by the machine, and dividing the number of intersections with the union of those two sets. I calculated the Jaccard index with the intent of providing some notion of partial agreement between the human coded and machine coded sets. Given that the model is fixed and not retrained dynamically, no cross-validation was performed. I tested taking the highest k sizes in the document for $k = 1, 2, 3$. I also calculated how many

# values	Non-none	# Equal	Equal proportion	Mean Jaccard index	Total
1	921	513	0.51	0.56	999
2	921	280	0.28	0.49	999
3	921	251	0.25	0.46	999

Table 2.12: Size accuracy

articles had the exact same size as reported by human coders. The results are presented in Table 2.12. The best results are when we report one size for each article: over half of the variables were equal to their hand coded values and the Jaccard index is at 0.55. Taking the top index is bound to create some false positives (for instance, an article reports a march that occurred in the past which trumped the size of the current event), and it may make sense instead to report the top two values instead of only one. Given that we don't see much of a decrease in the Jaccard index when two values are returned, this may be a better strategy for data production.

Overall, this is somewhat of a brute force method of extracting size, and there may be some more sophisticated methods for extracting size. One method would involve training an NER model to look explicitly for protest event size. Another could be to do a shallow parsing of sentences with a toolkit like Stanford CoreNLP and more explicitly identify subject-verb-object triads which look like protest size estimates. But for the time being, this represents a better solution than identifying numerical words in isolation.

Location: I used the CLIFF Entity Extractor created by the MIT Center for Civic Media[12] to identify location information. CLIFF identifies all sequences of text in the three classes used in the CoNLL 2003 shared task in named-entity recognition (Tjong Kim Sang and De Meulder, 2003): location, person, and organization, using the Stanford CoreNLP named-entity recognition classifier. Building on this, CLIFF enriches the location data by attempting to infer the "focus" of a news article. It combines this information with a geoparser

[12] `https://github.com/c4fcm/CLIFF`

called CLAVIN[13]. CLAVIN indexes the popular GeoNames gazetter, which contains several million geographical entries. CLIFF identifies organizations and provides comprehensive additional location information based on GeoNames, such as administrative districts, FIPS codes, latitude, and longitude. Accordingly, MPEDS returns up to four locations, their latitudes and longitudes from the source text, as well as the same information from the dateline.

In this paper, I offer no accuracy results for location because it is not clear what the best metric of this may be. While precision and recall metrics may make sense, the resolution provided by CLIFF may make it possible to estimate deviations from latitude and longitude from the human coded values. That itself may be more informative than just getting the actual text correct.

Social movement organizations. I use a custom-trained CoreNLP NER classifier to detect social movement organizations in the text. The out-of-the-box classifier for CoreNLP uses the training data from the CoNLL 2003 shared task. However, using this classifier yields a very high number of false positives, that is, a high number of organizations identified which are decidedly not social movement organizations. In an unaltered run of over many news texts, organizations include *senate, congress, supreme court, department of defense, walmart* and *monsanto*. These organizations are typically the target of protests and need to be distinguished from movement organizations.

I compared three different classifiers before settling on the newly-trained NER model. I first used a dictionary-based system in order to match social movement organizations. For that purpose, I integrated two organization dictionaries. The first is a cleaned set of organizations drawn from DoCA[14]. The second is a modified list drawn from the Encyclopedia of Organizations database (also called Associations Unlimited and thus abbreviated as AU)[15]. I collected all organization names in the Public Affairs, Labor, Social Welfare, and

[13] `https://clavin.bericotechnologies.com/`

[14] The Appendix in Wang and Soule (2012) report the process of matching and cleaning these data. Thanks to the authors for providing them.

[15] `http://find.galegroup.com/gdl/help/GDLeDirEAHelp.html`

Environmental categories at the time of writing. Further efforts to expand this database could retrieve the historical lists back to 1961. This task would take some effort to digitize paper records for years before 1995.

These systems, however, produced poor results, as noted in Table 2.13. F_1 for the Associated Unlimited database is 0.16, and it is 0.18 for DoCA. There are several possible hypotheses for this discrepancy. These databases tend to be US-centric, DoCA by design and AU by bias. DoCA excludes all events which do not take place in the US. For those categories which I collected from the AU database, 27,835 (21,568 regional and 6,267 national) organizations are based in the US compared to 6,481 based outside of the US. In addition, AU organizations had to be formally constituted, usually with a phone number and contact information, whereas many social movement organizations are briefly lived and not available for contact. Therefore it seemed to be more adequate to rely on an NER tool based on MPEDS-produced training data. In the future, however, DoCA could possibly be used as additional training data if their mentions can be matched to their source article texts.

I compare all the classifiers in Table 2.13 using random permutation cross-validation. I shuffled the dataset into a training and test set with a 90/10 split over three iterations and took the average of the metrics. The baseline classifier trained on CoNLL data yields an F-score of 0.32. There is high recall, given the inclusion of a wide variety of organizations, but the precision is 0.21, which is far too low for use in MPEDS. The classifier trained on MPEDS data, however, reports precision of 0.71, recall of 0.49, and F-score of 0.58, which is a dramatic improvement over the other solutions.

Future directions in extracting movement organizations could involve several more complicated steps. The solution of shallow parsing mentioned earlier in the size extraction subtask could be applied here. Furthermore, more dictionaries could be incorporated, such as the CAMEO and PETRARCH actor ontologies[16], and new dictionaries could be

[16] `https://github.com/openeventdata/petrarch/blob/master/petrarch/data/dictionaries/Phoenix.agents.txt` and `https://github.com/openeventdata/petrarch/blob/master/petrarch/`

Classifier	P	R	F1
au	0.55	0.10	0.16
doca	0.33	0.13	0.18
baseline	0.21	0.71	0.32
mpeds	0.72	0.49	0.58

Table 2.13: Social movement organization accuracy

developed with an eye towards intrastate protest and conflict. This paper also does not explore the possibly of ensemble methods, such as supplementing the MPEDS-trained NER classifier and allowing the dictionaries a (weighted) vote on each word in the article.

The open-ended coding task is the least developed part of the MPEDS system and at the same time one of the hardest. Extracting unstructured text from a set of news articles with heterogeneous origins is a difficult task. Size, location, and organizations are mentioned in a number of different ways and are embedded in a variety of different grammatical structures. However, given the evaluation presented here, it seems very hopeful that much more progress can be made and open-ended data can be extracted with a high level of accuracy. Several possible alternative solutions have been noted above. More are bound to be developed by engaging with the cutting edge in sub-areas of computer science and natural language processing, including new sequence classification models, and fast and accurate sentence parsing such as dependency parsing.

2.6 Discussion

The results presented here provide promise of rich event data for protest events with significantly less human intervention. They provide much more information than the sparse, dyadic datasets which are popular in the literatures in conflict forecasting and international relations. They also point to important improvements to current practice for generating event data for social movement scholarship. The advantage of time saved in

`data/dictionaries/Phoenix.MilNonState.actors.txt` generating data is the most important improvement. Once a relatively small amount of training data has been created, and classifiers have been created and validated, the actual data generation process involves the simple steps of apply the system to a corpus of news articles of interest. The training data produced by human coders by the MPEDS project amounted to some 20,000 unique articles, with nearly 3,000 positive news articles filtered for second pass coding. We also coded the articles more than one time to ensure reliability. With the MPEDS annotation interface, this work took less than two years with between 5 to 8 coders working at once. With these classifiers, we can produce protest event data from the whole of a large corpus, say the Gigaword corpus of 10 million newswire stories, in a matter of minutes. Compare this process to the decade-long task of producing the data for the Dynamics of Collective Action project, in which articles were coded only once and thus introduce more room for error.

The room for error is evident in the manual inspection of MPEDS's coding of DoCA data described above, in which 31 of 50 articles were "false false positives". DoCA is considered to be the state-of-the-art for protest event data in the US in the field of social movement research. It represents the largest US-based effort to collect a comprehensive account of protest events over its three and a half decades. This dataset of 21 thousand events, however, may not be complete and may misrepresent the true population of events which the *New York Times* covered over this time period. Over a dozen articles and at least one book have been produced using DoCA data. It would be a fruitful exercise to interrogate the results of this scholarship based on the data produced by MPEDS using DoCA's parameters.

That said, MPEDS as presented here is far from ideal. There are several areas which would be fertile ground for improvement. They are presented here from the most fundamentally omissions to the least.

Multiple events in a single article. While MPEDS excels at finding news articles mentioning social protest and conflict, it is only limited to addressing one event per article. A single news article may mention multiple events, for instance, a set of coordinated events occurring across different locations. This is where systems such as PETRARCH shine because they attempt to collect event data on the sentence level. Ideally, an a human-annotated training set would include information down to the paragraph and sentence-level in order to describe where information on the protest event can be found. Within the MPEDS project we have been able to collect these data into our hand coding system but have not yet implemented it into the classification process. If we are successful, it will allow us not only to identify multiple events in text, but to more accurately classify particular kinds of protest (either by form, issue, or target) with its appropriate textual components. Marakov et al. (2015) suggest using "anchor" words to identify protests at a sentence- or paragraph-level and merging them up into separable units. But their solution has not yielded results which are better than the results presented here.

Deduplication. Currently, MPEDS does not contain any mechanism for removing multiple mentions of a single event. This problem becomes more dire with the introduction of multiple news sources. This is a task within computer science known as record linkage[17]. Some of the fields produced by an MPEDS record may contribute to linking records together more accurately than others. But this is not addressed by the current version of the system.

Time-shifting. The system assumes that the event takes place on the same day (or at most, the day before) the reporting date. While this is often the case, it is not a guarantee. A possible solution would be to use a time annotation library such as SUTime (Chang and Manning, 2012) and perform an estimated time shift to more accurately report the date of the event.

Multiple languages. MPEDS is restricted to coding only English-language articles. This obviously introduces a Western (and moreover anglophone) bias into the events which are reported. However, cross-linguistic classification and information extraction is a large and growing area of natural language processing. Additionally, other than the size extraction

[17]`https://en.wikipedia.org/wiki/Record_linkage`

subtask and dictionaries available for organizations, producing new training data and generating new classifiers based on non-English languages is well within reason for the current machinery used to produce the MPEDS system.

Single closed-ended value. In the current version of MPEDS, the closed-ended variable can only take on one value at a time. While the MPEDS annotation interface allows for compound solutions (e.g. anti-war/peace and human rights), in practice these compounds are not robust enough across coders to make it into our final analysis. However, this is not a fundamental error. Instead of a single value, MPEDS could produce multiple confidence estimates for each variable. This may be skewed towards one category in which there is a high amount of confidence (e.g. abortion) but be useful for categories which have a significant amount of overlap (e.g. democratization and political corruption/malfeasance).

Temporal changes in language. Lastly, language surrounding protest can and does change over time. The history of movement scholarship itself teaches us that the rhetoric around movement activity once took the character of a "maddening crowd." Similarly, editorial and stylebook decisions can fundamentally change how a news source talks about an event. For instance, the virulent homophobia of former *New York Times* editor A. M. Rosenthal prevented the word *gay* from appearing in the Times (without scare quotes) until after his editorship ended, in 1986 (Elliot, 2014; Kaiser, 2012).

Most of these changes are not insurmountable, but they will take effort to extend the version of MPEDS presenting in this paper. All aspects of MPEDS are intended to be openly released and extendable by social and computational scientists. Currently, only the annotation interface of MPEDS has been released as open-source (`https://github.com/alexhanna/mpeds-coder`). However, by fall 2016 the whole MPEDS pipeline will be released as open-source software, most likely as a virtual machine package which encapsulates all of its necessary components.

2.7 Conclusion

This paper advances the state of protest event data generation in the social science using methods imported from computer science and statistics. Protest event data is a critical component in a number of subfields in sociology, political science, and other social sciences. Given the increasing availability of various electronic sources and the advances in natural language processing for large text sources, using automated methods is a natural move for social movements research and is essential for tasks like political instability monitoring, which requires real-time or near real-time data. The MPEDS project is innovative in combining machine learning approaches with deep substantive knowledge of the problems of identifying and coding collective events in news sources. MPEDS aims to be more efficient, replicable, and usable than the many of the current tools in social movements and political science. In addition, the data it produces will be rich enough for use in social movement research.

Even if many of the technical challenges presented here are resolved, many challenges will remain in the generation of new protest event data. First and foremost, the well-known biases of newspapers will not go away. Newspapers are businesses and as the institution of the news changes and consolidates towards less ownership, less diverse news may be a consequence. MPEDS cannot solve the consolidation and selection biases of news institutions, but it can allow for easier coding of new and more varied news sources. Additionally, reliance on news reports has forced researchers interested in generating event data to gather news articles from document databases such as Lexis-Nexis and Factiva. Although it may be easier to gather data in real-time for new events using web scrapers and RSS feeds, historical data is more difficult. These databases are typically only available to libraries which have purchased access, and even then they set up significant obstacles to downloading news articles at scale. The legal status of using news articles in event data research is thorny (Schrodt, 2014), and these databases understandably take an aggressive posture to limit access to the intellectual property they store. There may, however, be alternatives to this model of access for researchers. The Linguistic Data Consortium, for instance, provides free access to text repositories to member institutions. And the Open Event Data Alliance (`http://openeventdata.org`) is an attempt to develop similar infrastructure for event data researchers.

Institutional barriers notwithstanding, systems such as MPEDS have the potential to change the way we study social movements. They can allow us to ask questions and test hypotheses in ways which we couldn't have imagined in a prior era. For example, by generating protest event data in real-time, it may be possible to test hypotheses about mobilization in real-time through political forecasting, that is, making claims about events or trends in events which have not happened yet. Forecasting in social science can be used as a means of testing competing theories. Furthermore, advances in protest event data can allow for social movement theory to be updated in light of massive changes in communication and informational technologies, the constantly-evolving forms of social movement organizations and fields, and new movements against capitalism, authoritarianism, and racism, as well as contention stemming from right-wing nativist and anti-state movements. These measurement tools make possible the ground on which new social science theory can be built.

3 MEDIA ECOLOGY AND BACKLASH MOBILIZATION: BLACK POLITICAL REPRESSION AND #BLACKLIVESMATTER

> I keep coming back to what seems to me to be the most inhumane thing of all, the inhumane thing that happened before the rage began to rise, and before the backlash began to build, and before the cameras and television lights, and before the tear gas and the stun grenades and the chants and the prayers. I keep coming back to the one image that was there before the international event began, before it became a television show and a symbol in flames and something beyond what it was in the first place. I keep coming back to one simple moment, one ghastly fact. One image, from which all the other images have flowed.
>
> They left the body in the street.
>
> Dictators leave bodies in the street.
>
> Petty local satraps leave bodies in the street.
>
> Warlords leave bodies in the street.
> – Pierce (2014)

The death of Michael Brown in the summer of 2014 set off a firestorm of protest around the United States surrounding state violence against Black people. The simple but effective slogan, "Black Lives Matter" indicated the desire to be seen, heard, and recognized as humans, citizens, and social agents. Brown's death at the hands of a police officer was not the first of its kind, but the reaction of the Black residents of Ferguson grew from conditions not unlike those of Black people living in other places in the United States – a dense suburb or neighborhood of a large metropolis, a largely Black population with a large White police force, the housing, labor, and other economic crises of the Great Recession which have disproportionately harmed Black populations, and the carceral state exercising its power to police, surveil, and engage in mass incarceration.

Ferguson ignited a national movement against police brutality, state violence, and White supremacy. For a month after Brown's death, broadcast media and cable news ran information not only about riots and protests in the city of Ferguson, but also around the country when another unarmed Black individual was killed by police. Many narratives about the rise of the movement surrounded the use of the hashtag "#Blacklivesmatter" as a mobilizing call for other people to join in the movement. The hashtag signaled a social media savvy of its participants and echoed many of the images of the "Twitter revolution" which developed during the Arab Spring and the Occupy movement.

Movement scholars, however, know that the interactions of movements and the media are complicated and contextual. Social media does not create movements more than the *New York Times* does. These media have differential effects on movement activity and on attention given to movement claims. Additionally, different types of media interact with each other, some of which are supplementing the former activities of the others.

In this paper, I focus on the different levels of media attention to the Black Lives Matter movement. I discuss and expand on the concept, first espoused by Oliver (2008), that the modern policing of Black communities constitutes a form of political repression. Responses to extreme repression result in instances of backlash mobilization. I then develop a model of media ecology which operates after the repression event. I then use news sources and several computational methods to retrieve and process data around attention to the original repression event and the protest which follows it. Lastly, I evaluate the different levels of the media ecology model in a quantitative analysis.

3.1 Social Movements and News Ecologies

The interaction of movements and media operates through many different mechanisms, with various intentions of both movement and media actors. Movement actors hold public claims-making events in order to draw attention to an issue, to put pressure on public officials, to obstruct the daily operation of government or industry operations, and for a host of other reasons. Repertoires of contention change across time, as new cultural scripts of action become available with changing structural conditions (Tilly, 1978).

The self-conscious attempt to gain media coverage is a major preoccupation of many if not most modern movement organizations. Movements vie for media attention in order to enter the public sphere and affect popular discourses (Gamson and Modigliani, 1989) and to engage in more traditional "agenda-setting" (McCombs and Shaw, 1972), with the ends of forcing the hand of state and/or private actors. Within the US, this has become the sole occupation of many movement organizations, often to their detriment (Gitlin, 1980; Sobieraj, 2011).

The decision processes of what movement activity gets covered by news media and what does not are multilayered and dynamic, however. And with the advent modern social media technologies, this process has become more complicated. This section aims to highlight the interaction between 1) movements and broadcast media, and 2) how the introduction of new and social media has altered the relationship between movements and media. In a word, the selection processes of media attention are based in institutional and event-based factors. The introduction of social media has shifted some of the power to influence selection processes to actors, but much of the system is still in the hands of broadcast media and media elites. However, the nature of media elites has changed and includes social media and old media celebrities. The next section outlines those processes.

Media Coverage and Protest Events

Media coverage of movements by broadcast outlets, such as newspapers, television, and more recently, web-only publications, is shaped by a number of factors, rooted in movements and movement events themselves, news organizations, and the larger political landscape. Media coverage of movement events can be subdivided in those factors which affect whether the event will be covered at all (selection bias) and how the news media will end up covering the event (description bias) (Earl et al., 2004). Focusing on racially-based riots, broadcast media tends to be biased towards the intensity of the event and the distance from the news bureau or news office, event density, and city population (Snyder and Kelly, 1977; Myers and Caniglia, 2004). Broadcast media coverage is also sensitive to media issue attention cycles (McCarthy, McPhail and Smith, 1996). Furthermore, these patterns of bias do not seem to be stable over time (Ortiz et al., 2005). Coverage is also affected by legislative cycles and news hole effects, given there is a finite amount of space and a limited amount of resources that broadcast media can contribute towards coverage (Oliver and Maney, 2000). The interaction of movements and movement events is dynamic; movements and their claims enter into the public discourse only when there are opportune chances, or "discursive opportunities" (Ferree et al., 2002; Koopmans, 2004). Movements are aware of the nature of media selectivity and watch for those openings in which they can intervene and most forcefully make public claims.

More recently, scholars have engaged the question of not only whether a source mentions an event, but whether movement actors are able to voice their claims within the article, what has been called *standing* (Ferree et al., 2002, ch. 5). Sobieraj (2011), for instance, finds that, while their events may be covered at a higher rate, movement actors rarely achieve standing when trying to gain media attention at those events. Recent work by Amenta et al. (2015) have attempted to discern which elements may lead to the gaining of standing.

Thus far, the literature reviewed here has noted a particular type of interaction between broadcast media and movement participants. Movement participants engage in some event, and broadcast media is expected to pick it up and act as an amplifier for their message. Movement actors do engage in public relations-type strategies to gain that attention, including media trainings, practicing soundbites, and issuing press releases with their talking points (Sobieraj, 2011). It does not focus on the ability of movement actors themselves to be media producers and broadcasters. That ability has changed within the past decade, as the proliferation of new and social media has expanded at an incredible rate. While there are many who are pessimistic about the ability of these tools to act as effective broadcasters, (Hindman, 2008), a newer literature has developed which explores the possibility of movement actors being able to broadcast on their own and make their claims into the larger public sphere.

The literatures on the potential for social media to be a force for ground-up change in politics vacillates between the cyberoptimist to the cyberpessimist. Early proponents of these technologies suggested that their decentralized, networked nature would allow individuals to connect without any formal leadership, to avoid the high cost of central coordinating structures, and to reduces the barriers to entry in collective action, if not alleviate the problem of coordination in collective action itself (Benkler, 2006; Shirky, 2008; Castells, 2011). More recent work focusing on movement mobilization has been somewhat more measured in their claims about the potential for digitally-enabled movements to alter the nature of mobilization and social movement organizations. Earl and Kimport (2011) notes the various affordances provided by Information and Communication Technologies (ICTs), suggesting a continuum in which affordances are used sparingly to instances in which they are heavily relied upon. Bennett and Segerberg (2012) have noted the shifting nature of organizational dynamics in digitally-enabled movements, also suggesting a continuum in which networked groups of people use organizations to different effects, ranging from the self-organizing network to the organizationally brokered network. Similarly, Karpf (2012) details different types of advocacy groups and how they operate vis-a-vis larger political party and organizational structures. Some of these organizations have the possibility of disrupting the daily operation of political party machines and more forcefully making movement claims.

While organizations are critical to coordination of movement activities, one critical component of their activity is training of activities and having a free space for discussion of movement goals, tactics, targets, and organizational structures without hashing out these discussions in the light of broadcast media or for other opponents to take advantage (Evans and Boyte, 1992). Much of the literature on these online free spaces makes references to the idea of counterpublics, which serve as smaller public spheres in which subordinated social groups can invent and create discourses which counter those in the larger public sphere, to regroup and reformulate tactics and plans to influence and gain voice in that sphere (Fraser, 1990). More recently, researchers have focused on the availability of social media tools such as Twitter to operate as a counterpublic in the open, for individuals to use hashtags (keywords beginning with the # character) to discuss shared issues and concerns. In the case of the nascent Black Lives Matters movement, Bonilla and Rosa (2015) notes how hashtags have both a "clerical and semiotic" meaning. In the former the hashtag gives users the technical ability to search of tweets of interest easier (given that allows one to see popular and real-time streams of tweets using that hashtag by clicking on it). In the semiotic sense, it gives significance to the tweet in context, that this is what the tweet is really *about*. Jackson and Foucault Welles (2015) note the counterpublic that formed around the #myNYPD hashtag, which had been intended as a project of the New York Police Department to build solidarity around their officers. The hashtag, however, was turned against the NYPD, as Twitter users – many of them African-American and people of color – used the hashtag to tweet pictures and messages about police brutality, surveillance, and harassment. In a way, these "networked counterpublics" operate in a way that is reflexive and engaging with members of the counterpublic, but also have the ability to reach above that public to other publics. Some are intended to influence the larger public sphere, while some may do so unintentionally. In order to understand digitally-enabled movements more fully, we need to situate these movements within the larger news ecologies in which they are embedded.

The New Media Ecology

The concept of media ecology has its roots in media theorists such as Marshall McLuhan and Neil Postman. However, the object of study when one discusses a media ecology is nebulous if one is attempting to gain an analytical handle on it. Scolari (2012) notes how the media ecology becomes a stand-in for many different things, the main two being 1) larger media environment and 2) treating media as a species. The former view examines media as a set of institutions which have concrete effects on political, economic, and cultural life. The latter treats kinds of media as different kinds of beings that have to work within a singular ecosystem. In the species model, media species grow and/or die-off in a similar fashion as the classic Darwinian evolutionary paradigm; mechanisms of variation and selection allow some organisms to flourish within particular niches or parts of the ecosystem, while some weaken and die off in the face of competition. Media evolve not only in the face of political, economic, and social contexts, but they coevolve with each other. The media species respond to each other and feed off each other in various arrangements, some symbiotic and mutually beneficial, but some also possibly parasitic.

Any interrogation of digitally-enabled movements necessitates a broader look beyond their use of new ICTs and social media themselves. We need to introduce the understanding of the larger media ecologies in which they are situated to understand the claims-making activities and strategies of movement actors. While these tools can track the digital traces of human behavior in near real-time, focusing solely on social media and social media data tells us a one-sided story. For one, we know that social media are far from demographically representative of the larger populations in which they are situated. Certain platforms skew younger, more economically well-off, and are overrepresented in terms of African-Americans and Hispanics (Perrin, 2015). In prior studies of elections, Twitter activity has heavily favored leftists and political outsiders, such as Obama in the 2012 election (Shah, Hanna, Bucy, Wells and Quevedo, 2015) and the Pirate Party in the 2009 German federal election (Jungherr, 2013). News institutions have also realized the degree to which they must evolve along with new social media or risk losing their niche space in the larger media ecology. A working paper from Reuters details how social media has compressed the timeframe of the usual news cycle, changed the dynamic of "who" breaks news and shifts the role of news workers to those of curators and verifiers (Newman, 2009). The report also details how journalists and columnists use social media without organizational directive, instead using them to create a personal brand and information stream. Furthermore, while these technologies are coevolving with each other, mainstream and broadcast coverage is still highly socially significant; these organizations are still the main news sources which news consumers will turn for confirmation and validation of news activity. Social media adds another layer upon which people can discuss and interact with broadcast content.

For digitally-enabled movements, then, social media activity is part and parcel of larger media strategies. The media training many movement organizations go through now must extend to learning how to effectively leveraging social media for distribution of narratives and those claims established within movement counterpublics. Organizations and political communities have been founded to this end, often in league with political party organizations and large unions. The New Organizing Institute, until it suspended its operations in 2015, served that type of organization. The Netroots Nation conference is a mass meeting of the "netroots", typically technically-savvy people with affiliation with the Democratic Party or other liberal affinity. More independent organizations and networks are being continually created, such as the Allied Media Project, which is more oriented towards lower-income activists, queer activists, and activists of color.

3.2 Repression, Police Killings, and Backlash

The study of political repression of movements has seen a flurry of research within the past decade (Earl, 2011, provides a comprehensive review). Tilly's definition of repression is "any action by another group which raises the contender's cost of collective action" (1978, p. 100). This definition is expansive, as it allows the repressor to be any type of actor, including the national state such as a federal government, local government organizations such as police departments, and private citizens and groups such as the KKK. Earl (2003, 2011) has noted how the study of repression has focused extensively on overt, coercive activity undertaken by state actors. However, repression can also be covert (such as by state institutions such as COINTELPRO infiltration of radical militant Black organizations) and be built into the socio-legal structure of the national and local context, what Earl (2003) calls "channeling" (such as restrictions on 501(c)(3) social movement organizations).

Even when it comes to overt state repression, the usual treatments typically ignore significant aspects of state actions which demobilize and limit whole social movement sectors and broad swaths of people. Oliver (2008) notes the significant silence of political repression studies on the operation of ordinary policing on crime control, especially around mass incarceration and policing of Black people in the United States. She notes that in the idyllic telling of the Civil Rights Movement, the narratives around Black riots of the late 1960s and early 1970s have been left out completely. Those riots grew as a means of Black resistance against desolate poverty and abuse of White police in specific cities. Murch (2012, 2015) notes that in the Watts riots in 1965, rioters fought against massive overcrowding, redlining, and police abuse in the greater Los Angeles area. The violence of Watts by Black participants was directed against White-owned property, while police action was the result of the majority of injuries and deaths. The Watts riots saw the first deployment of the LAPD SWAT team, establishing the growing militarization of police forces. The Black riots in this period gave rise to the increase of mass policing and the establishment of the carceral state.

The analytical problem, Oliver argues, is that it is impossible to distinguish between dissent control and crime control. Dissent itself is often labeled as criminal behavior, especially in autocratic regimes and, in the United States, within Black communities. Murch (2012) explains that the social significance of Watts in the Los Angeles Black community took on a very different meaning than it did in popular (White) consciousness. The 1972 Wattstax rally, held on the 7th anniversary of the riots, celebrated the unrest and featured performances by prominent Black musicians and speeches by Black politicians. The problem with attempting to disentangle dissent and crime control stems from the fact that they have the same goal, the "maintenance of social order" and the protection of "unequal distributions of resources" (Oliver, 2008, p. 13). Furthermore, the policing that emerged as a result of the riots set the stage for the more intensive policing regime of Black populations, which has been maintained ever since. This has included the use of SWAT teams and the increasing militarization of police departments, as noted above, the establishment of the War on Drugs and the social order platform of Nixon and Reagan, and the rise of mass incarceration on low-level nonviolent offenses. Oliver also notes that repression operates through multiple mechanisms: deterrence, incapacitation, and surveillance. Mass incarceration operates on the level of incapacitation. Taking Black men out of their local communities means there can be no potential for organized resistance. Local and federal law enforcement agencies maintain a surveillance state element in many Black neighborhoods. On the local level, police remain on the lookout for Black individuals – mostly men – with outstanding warrants and unpaid fines, many of whom are parole and are at risk of violation and a return to prison or jail (Goffman, 2009). In its most extreme version, local law enforcement engages in large-scale surveillance in a similar manner to the federal government, such as the NYPD program of monitoring Muslims in the greater New York City metro area.

Police Killings and Backlash Mobilization

If policing regimes and mass incarceration represent instances of repression, then we also need to consider potential backlash to this repression. The question of whether repression increases or decreases mobilization is typically called the dissent-repression (or conflict-repression) nexus (e.g. Lichbach, 1987; Davenport, 2005). Prior findings from this literature have found that repression may have a number of effects: it may actually increase dissent, decrease dissent, result in a U-shaped response (i.e. decreasing dissent up to a certain amount of repression, then an increase again), or result in an inverted U-shaped response. I focus on what is called backlash (or backfire) mobilization, that is, the increase in mobilization spurred by increasing repression. Francisco (2004) analyzes the mobilization in the wake of intense state repressive, namely large massacres at protests. He describes the different mechanisms necessary for backlash mobilization. First, there needs to be enough information dissemination about the massacre to prompt a subsequent mobilization. This may take the form of an information cascade model (Lohmann, 1994), in which signals to mobilize meet actors' required thresholds to mobilization. Information dissemination also follows networked community of the repressions direct targets. Second, Francisco argues that movement leadership is required, not only in coordinating the backlash mobilization, but to adapt and adjust tactics to avoid further repression (Lichbach, 1987; Francisco, 1996). The largest risk backlash mobilizers encounter, Francisco contends, is being identified and subject to reprisal by the state.

It has been estimated by the Bureau of Justice Statistics that about 7,500 people have been killed by police in the US from 2003-09 and 2011 (Bureau of Justice Statistics, N.d.), but other estimates have put that number closer to 10,000 (Lum and Ball, 2015). In addition to the Bureau of Justice Statistics official numbers, several online projects have endeavored to collect information police killing, such as `killedbypolice.net` and the "The Counted" project created by *The Guardian*[1]. This also only represents killings which are reported in the media or in municipal or state records. Many of these killings go unreported or underreported in broadcast media sources. Police killings are heavily racialized and the victims are disproportionately Black men[2]. In 2015, *The Guardian* recorded 1,145 police killings. Black victims were killed at the highest rate of any racial group, at a rate of 7.22 per million (compared to 3.51 Hispanic/Latino and 2.94 White).

Within Oliver's framework, we can consider police killings as one of the most extreme forms of political repression within the non-authoritarian state context. Along with mass incarceration, police surveillance, police militarization, and the socio-legal bases of these

---

[1]`http://theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database`

[2]This is not to say mass policing does not also affect Black women and queer Black people. Black women are victims of many gendered and sexualized aspects of state violence, and are subject to indirect effects of mass incarceration and police killing in the domestic sphere (Goffman, 2009; Chatelain and Asoka, 2015). State violence is also a major factor in the lives in queer Black people. Transgender women of color, especially Black transgender women, constitute the largest percentage of those killed of all LGBTQ murders in the US. Their genders are frequently misreported in media and police reports. In 2015, over 20 transgender women of color were killed in the US, although none of these have been reported as police killings (Wikipedia, N.d.)

systems, police killings are a critical component of the carceral state. Police killing has both direct social and an indirect psychological effects on Black communities. Police repression has the direct effect of removing Black people from their local communities, of repressing their ability to amass political power and organization. It furthermore has the psychological effect of fear and anxiety about being a target of repression. These mechanisms are disproportionately felt by local networks composed of other Black people, typically, friends, family, and community members in highly segregated environments (Oliver, 2016). It is apparently by the many narratives of police killings of unarmed Black people that these killings seem to be indiscriminate, a result of "potential" crime control rather by any concrete action. John Crawford III was killed in a Walmart as he held a toy gun he picked up a store. Walter Scott was pulled over for a broken brake light and, after a physical altercation, was shot in the back by Officer Michael Slager after the officer withdrew his gun. These killings also appear to be state-sanctioned, as their perpetrators are rarely brought to jury trial or convicted. For instance, in the case of the killing of Michael Brown in Ferguson, Missouri, the St. Louis County prosecutor convened a grand jury, which decided to not indict Officer Darren Wilson. County and city prosecutors have been accused of holding conflicts of interest in these cases, given that they have close working relationships with the police departments they are charged with investigating.

Few movement scholars have examined police killings as an instance of political repression, and the protests that followed them as an example of backlash mobilization. As an important exception, Hess and Martin (2006) and Martin (2005) point to the beating of Rodney King in 1992 as an example of police repression and subsequent backlash mobilization. They examine the role of public opinion and the role of communicative structures in backlash mobilization. Returning to the model identified in Francisco (2004), there are some critical differences. In the first, Francisco makes the point that information environment is a critical component towards backlash mobilization. He makes the distinction that urban vs. rural environments, the fact that less dense networks or networks without higher connectivity to urban centers will not have sufficient communicative capacity. He also claims that the largest risk to backlash mobilization participants is being identified by the state. In the context of Black police killings, especially those which have taken place within the modern media environment, the status of these necessary claims may not hold, or rather, they are subject to more close interrogation.

3.3 #BlackLivesMatter and the media evolution of a movement

The Black Lives Matter (BLM) movement began in early 2012, in the wake of the Arab Spring and Occupy protests. 2011 was deemed by Time Magazine the "Year of the Protester" and many commentators labeled it as the global year of awakening. BLM was created in the wake of the killing of 17-year old Trayvon Martin in Sanford, Florida by George Zimmerman. As the well-trodden narrative is told, Martin, an African-American boy clad in a dark hoodie, was walking to the corner store to purchase iced tea and Skittles during the halftime of the NBA Finals. Zimmerman, a 28-year old Hispanic man, acting as a neighborhood watch coordinator, killed Martin, who he thought was involved in illicit activity. Zimmerman was eventually cleared of any wrongdoing. The social media attention around the Martin case did not begin immediately after his death. The killing was initially picked up on the crime beat by local newspapers, including the *Miami Herald* and the *Orlando Sentinel*. However, a concerted, traditional media campaign by Benjamin Crump, the lawyer for the Martin family, along with the publication and spread of a Change.org petition, led to more mainstream and social media attention (Graeff, Stempeck and Zuckerman, 2014). Social media spurred multiple protest and solidarity hashtags, such as #iamtrayvon and #millionhoodiemarch, which peaked about on Twitter about a month after the original shooting, when a special prosecutor was assigned to the case and President Obama made a statement about the case (Shah, Culver, Hanna, Macafee and

Yang, 2015).

The Martin case spurred the involvement and growth of several new movement organizations, created mainly by Black youth. The #Blacklivesmatter hashtag and Black Lives Matter social movement organization was created by three queer Black women, Alicia Garza, Patrisse Cullors, and Opal Tometi. The organization was created in July 2013, after the acquittal of George Zimmerman (Garza, N.d.; Freelon, McIlwain and Clark, 2016), along with several other organizations, such as Black Youth Project 100. Zimmerman's acquittal pointed to specific policy recommendations, namely the removal of "Stand Your Ground" laws which allow individuals to use deadly force to defend themselves without any requirement to evade or retreat, as well as "Castle Doctrine" laws which allow property owners to exercise deadly force in the case of an intruder on that property. The enactment of such laws are typically looped into the larger category of laws which enacted as part of a racialized repression apparatus; decisions to acquit or not indict offenders become flash points of backlash.

Nearly a year later, in July 2014, while selling single cigarettes on the street in Staten Island, New York, 43-year old Eric Garner became embroiled in an argument with two NYPD officers, Daniel Pantaleo and Justin Damico. The interaction went on for nearly five minutes and was completely videotaped on a cell phone. Garner was put into a chokehold by one officer and was brought to the ground while repeating "I can't breathe" eleven times. In the wake of Garner's death, several smaller demonstrations took place in New York. According to local media reports, these protests heavily involved Al Sharpton and his National Action Network. However, even with the exchange caught on video and occurring in a metropolis like New York, Garner's case did not gain significant media attention right away.

Michael Brown's story became the watershed which narrowed media attention on the trend of police killings. While Trayvon Martin's death highlighted the racialized violence which is generated by Stand Your Grounds and Castle Doctrine laws, Michael Brown's death implicated the carceral state and state violence by police as part and parcel of Black political repression. On August 9, 2014, in Ferguson, Missouri, as 18-year-old Michael Brown walked home with a friend, they encountered Officer Darren Wilson, who stopped them in the street. After a brief struggle, Wilson shot Brown several times in the front of his body. Local residents in the neighborhood caught video of Brown's body, surrounded by police cars. His body lie in the middle of the street for over four hours, an action which elicited outrage from local residents and, for residents, highlighted the disdain which the Ferguson police department held for the city's Black residents. Subsequent memorials left at the site of the shooting were destroyed by police cars, or urinated upon by police dogs.

In the immediate wake of Ferguson, Black residents began to riot and protest his death. Nights of clashes and heated exchanges between residents and police were sustained for several nights afterward. Residents on the ground began to livetweet the riots, sharing dramatic video and images. While the #Blacklivesmatter hashtag was created in the wake of Trayvon Martin's death, it did not gain much traction until the unrest in Ferguson became national and international news. The hashtag and the phrase became a rallying cry around which Black political power and claims came to the forefront. In the longer term wake of Ferguson, other shootings of unarmed Black people by police began to enter into broadcast news and social media narratives. Echoes of Ferguson focused media attention on killings which occurred both before and after the event. The BLM movement itself became more organized and well-resourced – at the time of writing, there are at least 23 chapters of the BLM organization in the US, Canada, and Ghana.

3.4 A Media Ecology Model for Backlash Events

The interaction of movements and media is complicated and multi-layered. Part of the complexity involves the processes of news making, of what journalists and editors decide to put into the publication, of the organizational resources available for a given story. A discussion on these elements is beyond the scope of this article. Instead, I focus on the interactions between different levels of media, and how those media cover different types of events over time. Towards this end, I present a media ecology model for backlash events. Figure 3.1 conveys a rough sketch of this model.

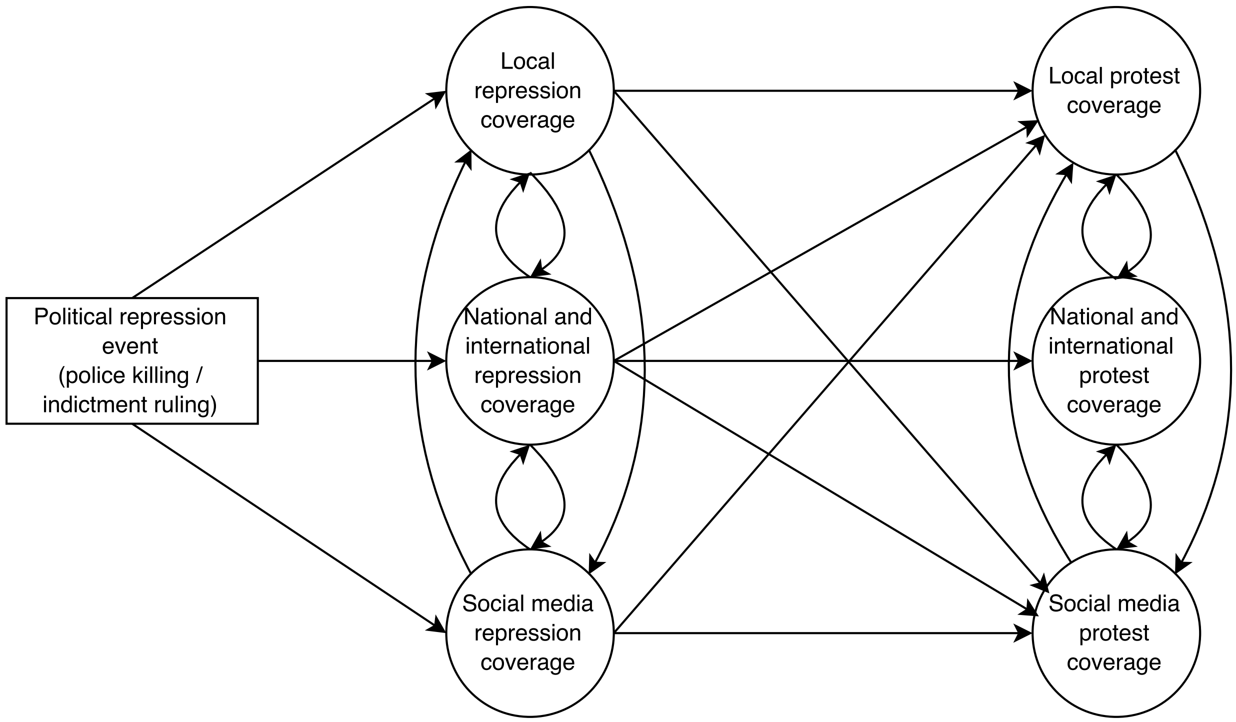


Figure 3.1: Graphical model of news ecologies of attention and protest event coverage.

At the beginning of the process, there is the repression event itself. This can take form of a police killing, or the announcement of a prosecutor's decision in whether to pursue an indictment of police or a jury trial. As noted above, the act of political repression in name of social order can take many different forms. Their manifestations as discrete events include police killings, mass arrests, and court decisions. The repression event itself has to be observable and able to be communicated to others.

After the repression event takes place, media entities pick up the story and broadcast it. It is at this point in which we reach the first stage of this media ecology, which within itself has multiple layers. The model presented here involves three of those layers: local broadcast news media, national and international broadcast news media, and social media.

This list of layers is far from exhaustive of all the possible layers within a modern media ecology. It could also include web-only news outlets (e.g. Vox, Politico) and partisan blogs and blog communities (e.g. Little Green Footballs and DailyKos). For the purposes of this paper, however, I focus on these three layers because each of them has the potential to play a different function within the media ecology. Local news media will typically cover local crime events, whether or not they can be linked to larger political movements. Local news coverage is more interested in the "local order" beat. They are also in closer proximity (both geographical and social) to informants who would talk to them about the event, including police spokespeople, neighborhood residents, and friends and family of the victim. National and international news media are typically associated with being *the* public sphere (Sobieraj, 2011). National and international news media are typically less concerned with the regular crime beat, of single incidents of police action, if they do not have a controversial, political, or intensive character. These sources will still cover isolated crime events, especially in their component "metro" sections (e.g. *New York Times* will typically cover these events, but further back in the paper)[3]. Lastly, social media can operate as a sphere in which the other layers interact, but also in which those people who are not news makers have a platform for information dissemination and political action. Social media users aim to share information with their local networks and larger imagined communities. These divergent purposes and expected audiences are different enough to represent significant coverage of the media ecology niche spaces.

What kinds of institutions exist within each of these layers? At the local level, news media includes primarily local daily newspapers and television station affiliates, but can also include local blogs and weekly newspapers. The national level typically represents the standard bearers of historical record (and of social movement research). These media include large national newspapers with daily circulations (*New York Times*, *Washington Post*, *USA TODAY*) and television news networks (CNN, ABC). At the international level, news

[3] In this paper, I treat national and international sources as a single layer, since they both serve similar purposes of being consider the larger public sphere and similar levels of audience reach.

wire services operate across countries and have established bureaus in many of the largest metropolises. These include the Associated Press, Agence-France Press, and Reuters. There are also international television networks, often directly or indirectly sponsored by state entities. These networks include Al Jazeera, the BBC, and Russia Today. Lastly, in the social media sphere, we can differentiate by platform (Facebook, Twitter, YouTube) but also by user. Many of the personalities and institutions in the broadcast media sphere operate with just as much activity in the social media sphere. Social media, however, does allow for "micro-celebrities" to emerge which are native to the social media sphere. Some of these individuals grow out of the context of a repression event itself.

This discussion of the operation of broadcast media elites preludes a larger view of the interactions of different layers in the media ecology after a repression event. These layers influence – and express influence upon – each other. At the local level, broadcast news institutions may be the first to pick up a repression event. Like in the Trayvon Martin case, repression events will often fall under the crime beat of the local paper, which, by themselves, are typically of little to no interest to national and international sources, and of possible interest to social media users in the area. When that repression event begins to generate backlash, it may then become of more interest to larger broadcast agencies and social media users beyond the immediate local community. At the national and international levels, stories and narratives which trace larger patterns of political abuse may have the power to influence local broadcast media to interrogate similar issues within that community. Social media users will often use national and international media as a touchstone to discuss local issues or issues salient to their group and community identities. Lastly, social media may have the power to break a story which had gotten no press at all or had been buried by local press. As Newman (2009) notes, updates about a breaking event may come from an average social media user or a nascent micro-celebrity. Broadcast media pick up details on the event, verify and curate that information, and include it in a larger article.

Those descriptions have focused mainly on coverage of the repression event itself. The event may or may not spur backlash mobilization. As Francisco (2004) notes, the backlash mobilization itself is a function of sufficient information dissemination towards potential protest participants. For the purposes of this paper, we can consider the protest event itself as an unobserved variable. What we have and use as a source for estimating a salient protest event variable are the news media which report on these events. Protest coverage itself is conditioned both on the larger media ecology in which is embedded, and the past coverage of the original repression event by all layers of the media ecology. A media source which heavily covered the initial repression event will be more attuned to cover subsequent protests in response to that event. A protest in one geographical location may gain local media, due to the national, international, and social media coverage of a separate repression event. For instance, solidarity events are one instance of this occurring (Almeida and Lichbach, 2003). National and international media may rely on local and social media sources to indicate the extent of protests occurring within a single movement. Lastly, social media users share articles from local, national, and international broadcast media to express their own solidarity (or disdain) for protests, to attempt to mobilize their own protest events, or to curate an online identity which matches their personally-held values or those of their friendship networks.

3.5 Data and Methods

My data come from two locations, newspapers and news wire sources (what I call collectively *broadcast sources*), and social media, namely Twitter. Newspaper data were collected from Lexis-Nexis using two methods. First, a broad search string was used to filter for protest events featuring African-American participants. Second, I queried for articles mentioning any of a number of keywords and victim names from a recent report on BLM (Freelon, McIlwain and Clark, 2016, p. 21), in addition to several others (Trayvon Martin, for instance). This keyword list includes more than a dozen Black high-profile police killing victims. Both the search string and list of victims are listed in the Appendix. The social media data come from the Twitter "gardenhose," a 10% sample of all of Twitter, stored in a database housed at the University of Wisconsin-Madison. The data are stored a distributed Hive database on a Hadoop cluster. From these data, I queried for any tweets mentioning those keywords in the victims list, plus the hashtagged version of all keywords (e.g. both "tony robinson" and #tonyrobinson, the #ferguson hashtag, and the #blacklivesmatter hashtag). Protest tweets were identified with the same broad search string. Both of these data sources cover the period from February 2012 to June 2015. This period includes the killing of Trayvon Martin in February 2012, Zimmerman's acquittal in July 2013, and protests around the deaths of Michael Brown, Eric Garner, and Freddie Gray in 2014 and 2015. I use the period after February 2014 to June 2015 for the main analysis below.

Although the search string strategy has the ability to gather a wide swath of protest events (Maney and Oliver, 2001), it does not narrow down the long list events, nor does it help us code anything specific about the event. To generate protest event data from the newspapers, I used a system developed by myself and collaborators called the Machine-learning Protest Event Data System, or MPEDS. MPEDS uses a mixture of machine learning and natural language processing tools to produce protest event data. Its classifiers are based on human-coded "training" data and achieves a high level of reliability in cross-validation procedures. MPEDS has three components – a haystack classifier, a set of closed-ended classifiers, and a set of open-ended extraction routines. The haystack classifier discerns whether an article mentions a protest or not. This, in and of itself, is a difficult task and has been where many other machine-aided event data systems have limited their focus (e.g. Croicu and Weidmann, 2015; Nardulli, Althaus and Hayes, 2015). MPEDS then classifies the form, claim, and target of the protest event from a list of predefined categories. The forms include common categories such as rallies, marches, riots, and labor actions. Claims include abortion, economic inequality, and, most salient for this case, racial and ethnic rights. Targets focus on the domestic government, universities, and private businesses. The last component is a set of open-ended routines which combined named-entity recognition, rule-based extraction, and gazetters. This allows us to extract the size of the protest, the organizations involved, and the locations of the event[4].

Data sources

The fruitfulness of this approach allows us to focus on several different levels of the media ecology at once. Of the large number of sources that Lexis-Nexis contains, I selected on those which had at least ten articles for larger media attention around police killing victims, and at least three for protest articles. The broadcast media attention dataset contains 21,325 articles from 90 news sources (80 local, 6 international, and 4 national), while the broadcast protest attention dataset contains 1,294 articles from 54 news sources (45 local, 6 international, and 3 national). The publications used are listed in the Appendix. Each of the broadcast dataset was cleaned to remove articles with duplicate titles, or articles which had the same lede sentence which occurred on the same date, in order to prevent counts which are inflation by syndicated content. The total Twitter dataset contains 8,199,690 tweets, while the protest Twitter dataset contains 281,717 tweets. Retweets have not been removed in order to retain a truer metric of social media user attention.

[4]This system is only suitable for analyzing newspaper and news wires articles. Therefore, I do not use it to identify protest-related tweets

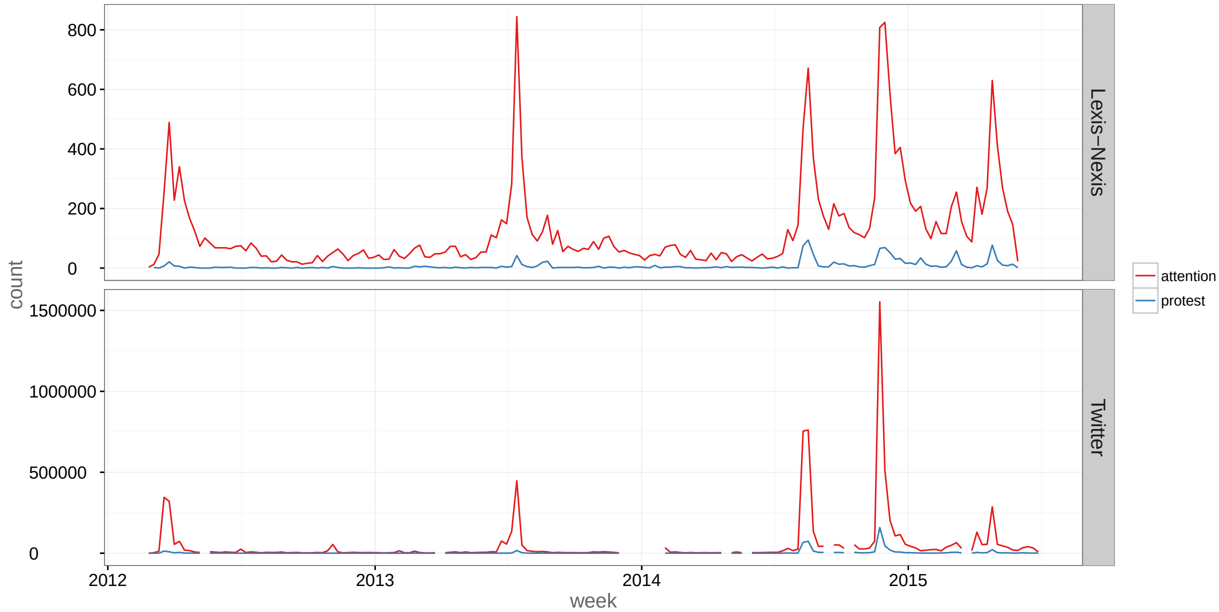


Figure 3.2: Attention to police victims compared to protest mentions in Lexis-Nexis and Twitter, Feb 2012 - June 2015. The Twitter dataset contains several periods of missing values due to technical issues.

Figure 3.2 compares total attention to repression and protest events over time for both data sources. The peaks in both graphs correspond to major repression events and responses to those events. The initial peak in early 2012 is the response to the Trayvon Martin killing, which, as mentioned above, comes several weeks after his death. The next major peak occurs in July 2013, when George Zimmerman is acquitted of charges. It's notable that this produces a smaller comparative peak in Twitter compared to broadcast sources. For broadcast sources, it is almost the highest point of attention for the data period. The next peak occurs days after the killing of Michael Brown in Ferguson. There seems to be scant attention paid to other killings occurring before that, including that of John Crawford and Eric Garner. The second-to-last peak is the announcement by the St. Louis County prosecutor that the grand jury will not indict Darren Wilson. The last peak represents the killing of Freddie Gray and the protest events which occurred after them.

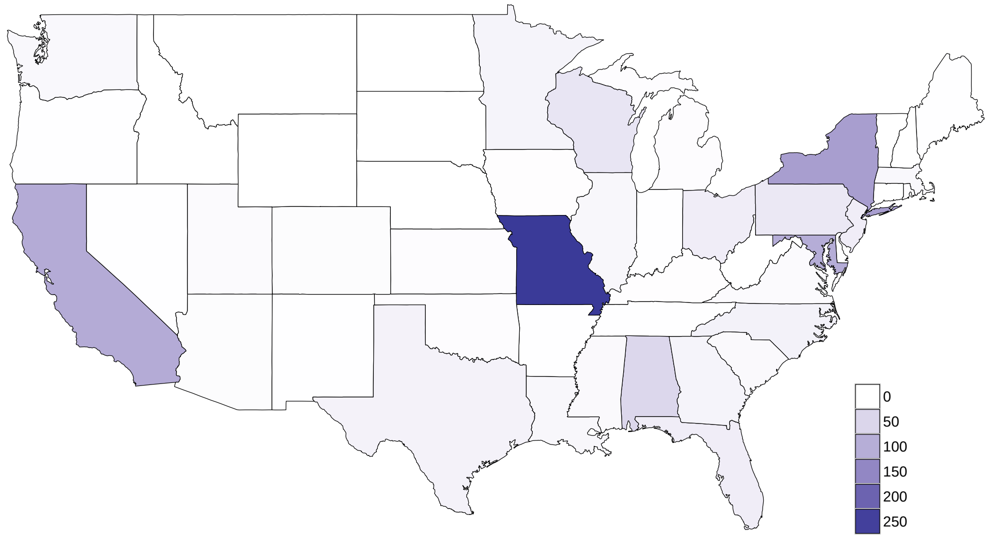


Figure 3.3: Map of Black protest events mentioned in Lexis-Nexis, February 2014 to June 2015

Figure 3.3 notes the number of protests in each state from February 2014 to June 2015. Not surprisingly, the largest number of events are reported in Missouri, typically from Ferguson and the larger St. Louis metro area. More protests occurred in Maryland, specifically in Baltimore and in response to the killing of Freddie Gray. Protests occur in large population centers of Los Angeles and New York. The protests reported in Alabama are a mix of reports on a march which commemorated the 1965 Selma march, and a misclassification of the original protest as occurring recently. Other protests stem from Sanford, Florida, Cleveland, Ohio (surrounding the deaths of Tamir Rice and Tanisha Anderson), and Madison, Wisconsin (surrounding the death of Tony Robinson).

Analytical strategy and research questions

Given the flurry of activity which usually takes places directly after a discrete, public repression event, my main unit of analysis is the repression event itself. I focus on 20 of the police killing victims on which I collected data, and use both the date in which police exercised repressive action and the date of an indictment decision of the responsible police officer(s)[5]. From each of these dates, I generate variables which measure broadcast and Twitter media and protest attention to these victims from the date until two weeks after the event. The counts are the number of articles and tweets which mention the victim for each day in the time period[6]. Of these dates, I model the counts of each media type the day after the repression event, using the counts of media on the day of the repression event as my independent variables. Because all variables are count data and they are overdispersed, I used a quasi-Poisson regression model for estimation.

Based on the media ecology and backlash protest model presented in Figure 3.1, I propose several research questions to test from these analyses. My first set of questions concerns the public attention media ecology. How do different pieces of broadcast and social media fit together? I ask the following three questions.

> *Research question 1 (RQ1)*: How do national/international (nat/int)[7] broadcast media, and social media affect coverage of repression events in local broadcast media?
>
> *RQ2*: How do local broadcast and social media affect coverage of repression events in nat/int broadcast media?
>
> *RQ3*: How do local and nat/int broadcast media affect coverage of repression events in social media?

Second, we wish to learn how media attention in the ecology impacts the coverage of protest events at different layers. In the model above, media attention and coverage of protests will influence how protests events are covered on the local, nat/int, and social media levels. Therefore, the next set of research questions focuses on protest events.

> *RQ4*: How do media around repression events, as well as protest coverage from nat/int broadcast and social media, affect coverage of protest events in local media?

[5]Three of these cases had no discernible indictment decision date and thus not used in the analysis.
[6]For tweets, I merge the name mention and the corresponding hashtag.
[7]I abbreviate national and international news this way in the rest of this paper for convenience.

RQ5: How do media around repression events, as well as protest coverage from local broadcast and social media, affect coverage of protest events in nat/int media?

RQ6: How do media around repression events, as well as protest coverage from local and nat/int broadcast media affect coverage of protest events in social media?

3.6 Results

Figure 3.4 presents a histogram of logged counts for each data source and type of tweet or articles. The unit of analysis are the daily counts for each day after the event. For all of the conditions and data sources, there is no coverage for that victim for most of the two weeks after the repression event. Twitter repression attention seems to be more uniformly distributed, but Twitter protest still has a significant number of zeros. There is much less volume overall for protest messages, as Figure 3.2 indicates above. There are several outliers as well, especially with regard to Twitter. This is due to the significant attention given to Michael Brown after his killing and Wilson's non-indictment.

Figure 3.5 displays the logged counts of media instances for the two weeks after the event. The gray lines represent the individual cases and the dark red line is the mean. In the social media case, for both the attention and protest variables, the line peaks in the first day after the repression event and then slowly goes down for the rest of the period. The broadcast news graph presents the differences between local news and nat/int news. Both the attention and protest variables take a different shape than the Twitter post-event trend. There is certainly a bump in the first day after the event, but there is also an increase in attention in the second day after the event. Nearly a week after the repression event, there's another bump in media and protest attention. This may indicate the influence of the news cycle which doesn't appear in the Twitter case. There is also a good deal of variability between the different cases.

Research questions 1 through 3 focus on media attention on the repression event itself.

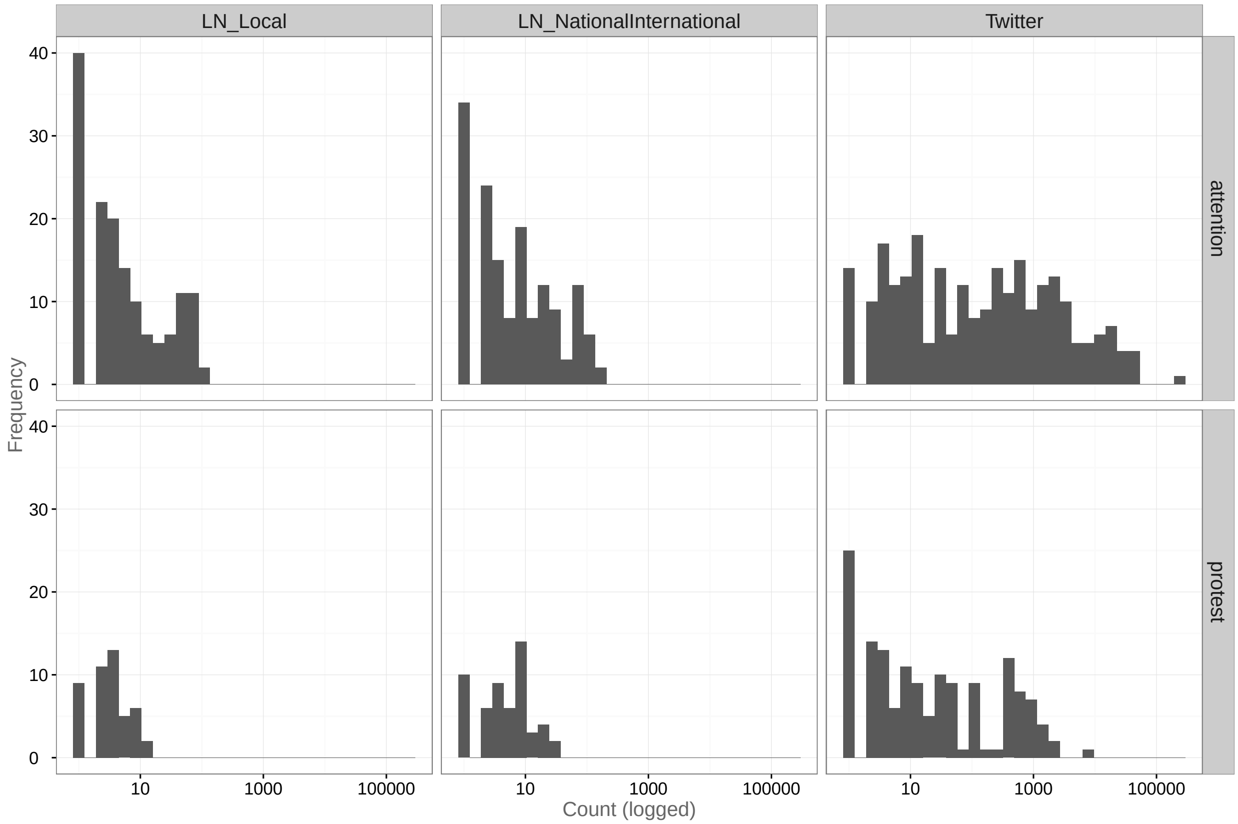
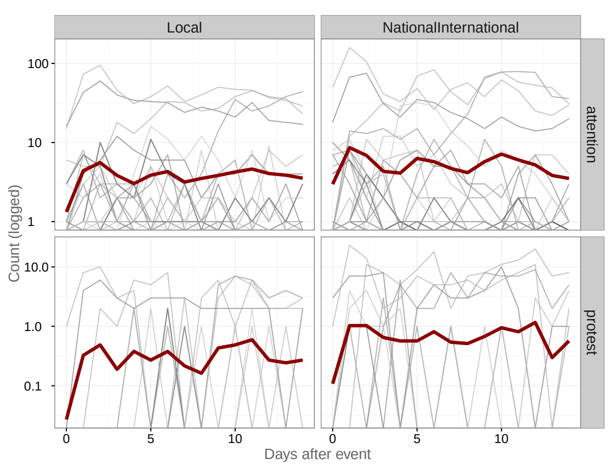
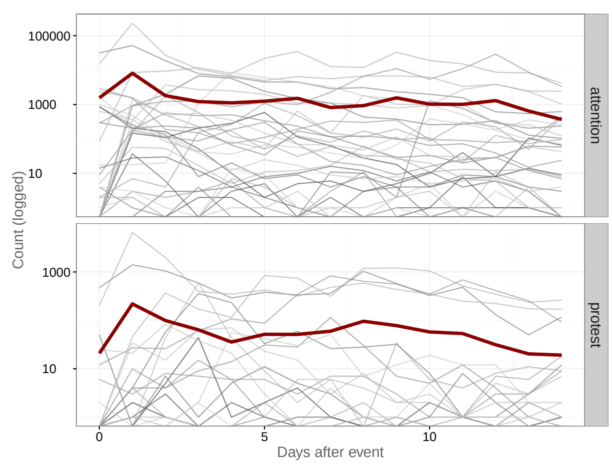


Figure 3.4: Histogram of attention and protest for each layer.



(a) Broadcast news articles, by geographical scope



(b) Twitter

Figure 3.5: Count of attention and protest messages (logged), two weeks after repression event. Red line is the mean.

Table 3.1 presents the results of the regression analysis of the attention variables. In the case of local broadcast media (models 1–4), both Twitter and nat/int media attention are seen to be positive and statistically significant predictors of local media attention. Past local attention also positively affects future local attention. In the saturated model (4), however, only prior local attention positively predicts future local attention, and is statistically significant ($p < 0.05$). Similar results hold for nat/int media attention of the repression events (models 5–8). Both local and social media positively predicts nat/int media, as well as prior nat/int attention. However, in the full model, local media has a positive statistically significant effect on nat/int repression coverage. Lastly, in Twitter coverage of the repression event (models 9–12), again we see positive and significant bivariate relationships. In the full model, prior Twitter attention and nat/int attention has a slightly negative effect on future Twitter coverage. Most surprisingly, local media attention to the repression event has the largest effect of any (0.61) and is statistically significant ($p < 0.001$).

These models taken together would seem to indicate that local media remains critical during repression events, and that it is an important predictor for national and international broadcast and social media diffusion (RQ2, RQ3). However, the inverse cannot be said of nat/int media and social media, given that they had no effect on local media (RQ1).

Table 3.1: quasi-Poisson regression analysis of attention variables

	Local-attn$_{t=1}$				Nat/Int-attn$_{t=1}$				Twitter$_{t=1}$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Twitter-attn$_{t=0}$	0.0001***			-0.0000	0.0001***			-0.0000			0.0001***	-0.0001^{***}
Nat/Int-attn$_{t=0}$		0.08***		0.01			0.08***	0.01		0.11***		-0.002
Local-attn$_{t=0}$			0.26***	0.30*		0.25***		0.29*	0.36***			0.61***
Constant	0.77	0.56	0.07	-0.01	1.51***	0.79**	1.16***	0.70*	6.17***	7.03***	8.16***	5.34***
N	37	37	37	37	37	37	37	37	37	37	37	37

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 3.2: quasi-Poisson regression analysis of protest variables

	Local-prtst$_{t=1}$			Nat/Int-prtst$_{t=1}$			Twitter-prtst$_{t=1}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Local-attn$_{t=0}$	0.23		0.00	0.16		0.43	0.48**		0.49***
Twitter-attn$_{t=0}$	0.001***		-0.00	0.0000		-0.0001	-0.0001		0.0004
Nat/Int-attn$_{t=0}$	0.35***		0.00	0.05		0.26	0.03		-0.38
Twitter-prtst$_{t=0}$		-0.00	-0.00		0.06**	0.02		0.02	0.10
Local-prtst$_{t=0}$		19.15***	19.15***		1.24	-11.81		4.52***	11.68
Nat/Int-prtst$_{t=0}$		9.23**	9.23		-8.18^{*}	-5.10		-1.31	-17.31
Constant	-28.45^{***}	-26.30^{***}	-26.30^{***}	-1.74^{**}	-2.00^{**}	-3.11^{*}	0.90	1.44**	0.93
N	37	37	37	37	37	37	37	37	37

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Turning now to coverage of protest events, I constructed three models per type of media, presented in Table 3.2. The first model focuses on prior attention to the repression event itself. The second focuses on prior coverage of protest events. And the third model combines the two. In the case of local protest coverage (models 1–3), prior local, nat/int, and Twitter attention to the repression event have a moderate effect, and both Twitter and nat/int coverage are statistically significant. The same effect holds for prior protest coverage, with prior local news having a large effect, and nat/int coverage having less of one. In the saturated local protest attention model, only prior local protest coverage remains significant.

For nat/int coverage (models 4–6), prior local and nat/int repression attention have moderate effects but do not achieve statistical significance. However, in the prior protest coverage block, Twitter and nat/int media sources are statistically significant. The largest effect is prior nat/int protest event coverage. However, it has the opposite effect from what we'd expect – prior coverage seems to depress future nat/int coverage. In the saturated nat/int protest attention model, the sign switches on prior local protest coverage and becomes a largely negative predictor. All variables lose statistical significance.

In the final set of models focusing on Twitter protest coverage (7–9), prior attention is statistically significant and positive for local repression attention. Local attention has the largest effect, similar to the result achieved in model 12 of the Table 3.1. In the prior protest attention block, local attention has a large and statistically significant effect. In the saturated Twitter protest model, local repression attention remains positive and grows slightly. The effect of prior local protest in the saturated model more than doubles (from 4.52 to 11.68), and prior nat/int protest coverage has a large negative effect (-17.31). However, neither of these large effects achieve statistical significance.

These results again suggest the importance of local media in coverage of both original repression events and the subsequent protest events (RQ5, RQ6). However, it is a curious result that prior protest coverage by nat/int sources seems to depress future nat/int and

Twitter protest coverage. It's not clear whether this is a data issue, a result of model selection, or represents some substantive effect without an obvious mechanism for explanation.

3.7 Discussion

These results indicate that there are intricate interaction effects between the different layers of the media ecology. In both Tables 3.1 and 3.2, prior local attention to repression and protest events is a consistent predictor for local, nat/int, and Twitter repression coverage. It is still a strong predictor of local and Twitter protest coverage, but not nat/int protest coverage. This would seem to indicate that the place of local broadcast news is an important component in how the media ecology treats Black political repression and backlash events. This would seem to make sense, given that local news will typically report on these events more consistently and as part of the regular news making work of the local newsroom. But this result has implications for how we study Black political repression and its backlash, and social media in protest.

First, because Black political repression of the kind characterized by Oliver (2008) is looped into the category of regular policing and social control, it means that the institutions who are most wont to cover these stories will be local broadcast media. Given that reporters are on the common crime beat, there is a more significant eye towards events which are related to that beat, including followup protest events. This may actually be a good thing for social movement research, since this research almost exclusively uses news sources to study protest events. We know that there's a number of selection processes which influence what actually ends up in the newspaper, and that it has been argued that these processes are not stable across time. But if there is consistent attention to repression events as part of regular newspaper practice, then it may be the case that the selection processes which guide what events enter into broadcast news are more consistent across time. This is not to say that protesters and movement activists will be legitimized or given standing by journalists, however. There is a non-trivial amount of research which indicates that they may be maligned (e.g. Davenport, 2010). However, it may mean that local news sources allow a more consistent data source for gather event data on backlash mobilization events.

Second, the promise that social media is a critical component for the publication of information about Black repression events is probably too optimistic. Social media has its virtues, for sure, but there is still the necessity of local news organizations to sift through the haystack of Twitter feeds, Facebook statuses, and Vine videos to construct narratives around the repression events. Because of the limitations of this dataset, we cannot retrieve information on what time of the day that local and nat/int stories were released. We do have that information with Twitter, however. Many events unfold in the narrative of a single day. For instance, *Mother Jones* ran an article entitled "10 Hours in Ferguson: A Visual Timeline of Michael Brown's Death and Its Aftermath"[8] which details the narrative of events on August 9, 2014 around Michael Brown's death. The article embeds contemporaneous social media messages about the event – most significantly, videos and images. But critically, it also embeds reports by local news and news wire photographers. The case of Ferguson is an outlier in terms of repression events, but it would seem that the interaction of social media, local news, and nat/int news provide different functions and identify specific niches within the larger media ecology.

This analysis has several issues which could be explored in future work. First, the sole focus on the day of the repression event and the day after limits the potential slower effects. Protests may have a longer onset time, and news cycles influence the rate of news attention to a particular case. Using a time-series analysis would be appropriate here, given that it could take into account multiple time steps instead of a single one. Relatedly, the sample size is rather small. Few cases of police killing become large national stories, and most do not produce unrest on the scale of Ferguson. Given this fact, focusing on cases of isolated instances of police killing may be a somewhat limited research approach.

[8] http://www.motherjones.com/politics/2014/08/timeline-michael-brown-shooting-ferguson

Second, it would be helpful to disentangle the different local sources and to control for the distance any given local source is from the location of the repression event. Local news is certainly not heterogeneous, and the argument about local sources assumes that these sources are focusing on the repression event in question.

In terms of social media data used here, there are two issues regarding the selection of cases. Social media is an ecosystem upon itself. Facebook is often attributed for use with mobilizing activities (Hanna, 2013), while Vine and YouTube may be more useful in presenting the visual face of repressive events. Twitter data, however, is readily available, and can be queried easily for a set of keywords if one has collected the data in real-time or purchased it from a data provider. But the point is well-taken that Twitter should not be the model organism (*drosophila melanogaster*) for social media research (Tufekci, 2014). The keyword strategy is sound here because I am more interested in Twitter attention than actual mobilization activity. Twitter is also an appropriate platform to study, given that BLM organizers themselves created the movement around a hashtag, which is a Twitter-ism that has entered into nearly all modern social media practice.

3.8 Conclusion

This paper is an investigation into the media surrounding the Black Lives Matter movement. It expands on the notion that regular crime control and policing serves as form of Black political repression. Police killings are intensive expressions of that repression, as well as the non-indictment of the police officers which are involved in these killings. Protests which occur after these events are forms of backlash mobilization. This is unique insofar as discussions of repression and backlash have usually focused on authoritarian states or states with militant and violent factions. This paper also develops a model of the operation of media ecology in response to a repression event, considering media attention processes at large and attention to protest events specifically. It uses computational methods to generate metrics of news media attention, including the Machine-Learning Protest Event Data System. I find that local media attention is critical for both national/international attention and social media attention to both the original repression event and protest events.

This paper is innovative in both theoretical and methodological ways. First, it takes seriously the claim by Oliver (2008) that policing of Black communities is a form of political repression, and thus makes the next theoretical step of connecting protests in response to killings as backlash mobilizations. Furthermore, it attempts to develop a model for the operation of the media ecology during repression and backlash mobilization events. Methodologically, it uses a new machine-learning system for the generation of protest event data. It also relies on a large number of local sources and news wires, as opposed to the single newspaper focus of much protest event analysis research.

Future directions for this research should look more into the other practices of Black political repression by policing, including the elements of surveillance and incapacitation, both of which are by definition more hidden and do not enter into broadcast media often. In what ways can we research mobilization against mass incarceration and police surveillance? Furthermore, future work should be attentive to the discourses within articles on repression events. Victims of intense repression often go through a process of delegitimization and defamation (Hess and Martin, 2006; Martin, 2005). Do protesters within backlash mobilization receive the same treatment? Recent work on the Egyptian revolution suggests that different layers of the media ecology write about protesters in very different ways (Hamdy and Gomaa, 2012). Is this the case in different levels of the American media ecology during the course of backlash mobilization to Black repression events?

Future work should also focus on the longer temporal cycles of the Black Lives Matter movement, and how the events of Ferguson operated as a critical event during the cycle. While Ferguson was a catalyzing incident, the current cycle of protest around shooting deaths of unarmed Black individuals began as far back as 2012, and further echoes the Watts riots and the riots which followed the beating of Rodney King (Murch, 2015). Understanding the determinants of media attention and protest throughout the cycle is an important endeavor. As the leaders of BLM have say often “this is a movement, not a moment.”

4 CONCLUSION

This dissertation presents work on the use and development of protest event data. It analyzes the lineages of political event data, outlines the development of the MPEDS system, and applies it to a modern digitally-enabled movement. To summarize the points of this dissertation succinctly:

1. Two lineages of political event data developed in the past several decades. One was focused on the explanation of interstate conflict and internal channels to state stability. The other was concerned with the emergence of social movements and their consequences. In a quantitative analysis of the Civil Rights Movement, data used from the first lineage is unsuitable for use in the latter lineage. Automated methods may find ways to unite the two lineages once again, however.

2. It is possible to develop and create new automated tools for protest event analysis. The development of a new system, the Machine-learning Protest Event Data System (MPEDS), integrates tools from statistics and machine learning in order to create protest event data with little to no human intervention. The system has three parts – one which discerns articles mentioning protest from those which do not; a set of classifiers which classify form, claim, and target of the protest; and a set of tools which extract size, location, and social movement organizations involved. There are important differences in how well heterogeneous news sources can be used for training machine learning algorithms.

3. MPEDS is then applied to look at the media ecology of backlash mobilization of the Black Lives Matter movement. Mass incarceration and heavy policing of Black communities constitutes a form of Black political repression, and police killings of unarmed Black people constitutes one of the highest forms of this political repression. Responses to these repression events can be considered backlash mobilization. The

functioning of the media ecology following these events has multiple layers at the local, national, international, and social media levels. In the quantitative analysis, local media are found to be important to the coverage of the repression event and the protests which follow it.

This dissertation is an initial foray into the development of an automated system for the generation of protest event data. There are several teams within sociology and political science pursuing the same task, and I have begun some collaborations with some of them (e.g. the POLCON team in Zurich and Florence). As mentioned in chapter 2, there are many improvements that need to be main to the MPEDS system as presented here. First, the system needs to consider multiple events within a single article and be able to discern between their different locations, actors, and claims. Second, location information must be validated for each of the events, and we should be able to retrieve multiple locations per one article. Third, MPEDS does not do any deduplication of events, but record linking is a promising way forward here. Fourth, time shifting articles by mentions of relative times (e.g. yesterday) is necessary and also technically feasible. Fifth, we should aspire to create coding systems which are able to handle multiple languages, not just English. Finally, we should be attentive to temporal changes in language (what is called semantic shift in linguistics).

The wonderful thing about political event data and automated methods is that researchers who work in this tradition tend to highly value openness and collaboration. Phil Schrodt's originally TABARI code was written in PASCAL and then C++, and he has been posting his code to the web well before it was the norm for programmers, not to mention social scientists. The new PETRARCH system has been posted to GitHub and is a collaboration between multiple institutions. One takeaway from this is that social movement scholars should also be committed to this sense of openness in tools and their dissemination.

Extensions to MPEDS could be made in many ways with a bit of coding work and generation of new training data. Datasets specific to more general types of events, like crime reports or police killings, could possibly be created with a similar framework. The groundwork MPEDS lays could lead to alternative ways of defining and creating any sort of "event-ful" sociology. Science is necessarily a cumulative process, and I hope this foray into tool-building provides many bases upon which sociologists, information scientists, political scientists, and computer scientists can built.

A APPENDICES

A.1 Software – Chapter 1

All computation and visualization in this paper was performed with the R statistical package, version 3.2.3 (R Core Team, 2015). Graphs were produced with `ggplot2`, version 2.1.0 (Wickham, 2009). City and state locations in SPEED were obtained through reverse geocoding latitude and longitude with `ggmap`, version 2.6.1 (Kahle and Wickham, 2013). Maps were drawn with `maps`, version 3.1.0 (Brownrigg, 2016).

A.2 Definition of an MPEDS event

For this project, a protest event should be coded "yes" if there is sufficient information to identify a discrete protest event that occurred in the past or is being planned for the future. A protest mention should include information about the time and location of the event and the actors or claims involved. More specifically a protest mention should be an event that involves:

1. Location of the protest and date that the event occurred
   a) Protest can be occurring, have occurred in the past, or is being planned for the future
2. A claims-making or grievance expression. There is a message that is attempting to influence social change, policies, or actions by major institutions. The message can be advocating for or opposing policies advocated by others.
3. Some of the actors must not be government officials or institutional leaders like university officials or business owners.
   a) An actor does an action or makes a claim; merely being in the audience does not count.
   b) Actions by leaders of organizations whose purpose is to advocate for social change goals ARE protest actions. This is the difference between, say, the Chancellor of the University and the head of the National Organization for Women.
   c) We will leave in actions initiated by unions if they involve claims-making or expression of grievance.

List of acceptable protest events:

- Rally, strike, protest, march, demonstration, vigil, picketing, leafleting, occupation or sit-in, blockade/slowdown/disruption, strike/walkout/lockout, civil disobedience, boycott, riot, property damage, symbolic display, press conference or news conference by a non-governmental organization (NOT an elected official, NOT government agency, NOT candidate for office), commemoration, counter-protest, walk-out, sit-in, ballot events, online protest (has to meet same criteria as offline protest).

- Routine events by social movements groups (e.g. gay pride parade).
- It is likely that there will be other forms of protest (e.g. reading a banned book, chaining oneself to a desk, carrying a mattress). These can be included but must be connected to larger claims-making movements.

List of NOT acceptable protest events:

- Individual grievances, crimes, and public displays of anger that are not expressive of either social goals or the shared grievances of a social category/group.
- Isolated crimes, lawsuits, court cases, policy decisions that do not include expressions of a broader group or collective.
- Political campaign events.
- Violence or vandalism unrelated to a broader claim or issue; the act of violence is primary form of protest (e.g. bombing of Sterling Hall).
- Press conference or news conference by a governmental organization (elected official, government agency, candidate for office).
- "Routine" politics, political campaigns, speeches by elected officials or those seeking electoral office.
- Interviews and speeches by celebrities.
- Simple charity activity (collect money, no policy demands).
- Entirely personal actions (e.g. crime, personal expression), self-expression (purpose of gathering is entirely inward, not outward). A party or religious ritual as an end in itself is NOT a protest event, but these forms could be protest events if the stated purpose is to make a claim, express a grievance, or advocate for/against some policy change.
- Acts of terrorism – a rough definition of terrorism as an event aimed at intentionally destroying large amounts of property or injuring people.
- Selma rule: exclude coverage of historical protests that are not happening contemporaneously; we only want articles that discuss protest events that are directly relevant at the time of writing, at least within a year of the reporting date.

A.3 Search string for MPEDS articles

We use a search string to collect data from Lexis-Nexis and to filter out articles in the New York Times Annotated Corpus and the Annotated Gigaword collection. Our intent with this string is full recall without regard to precision. Therefore the string is as inclusive as possible. This was constructed by referring to existing movement search research (Maney and Oliver, 2001; Strawn, 2010). In our testing, we were able to retrieve all DoCA articles over periods of highest event density.

boycott* OR press conference OR news conference OR protest* OR strik* OR rally OR ralli* OR riot* OR sit-in OR occupation OR mobiliz* OR blockage OR demonstrat* OR marchi* OR marche* NOT protestant*

A.4 Software – Chapter 2

Article data were processed using Python version 2.7.10, `pandas` version 0.17.1 (McKinney, 2010) and `numpy` version 1.10.4 (van der Walt, Colbert and Varoquaux, 2011). MPEDS was built with scikit-learn version 0.17 (Pedregosa et al., 2011), Stanford NER v3.6.0 (Finkel, Grenager and Manning, 2005), and CLIFF (`https://github.com/c4fcm/CLIFF`). Visualization and correlational analysis was performed with the R statistical package, version 3.2.3 (R Core Team, 2015). Graphs were produced with `ggplot2`, version 2.1.0 (Wickham, 2009).

A.5 Software – Chapter 3

Lexis-Nexis and Twitter data were processed using Python version 2.7.10, `pandas` version 0.17.1 (McKinney, 2010) and `numpy` version 1.10.4 (van der Walt, Colbert and Varoquaux, 2011). Twitter data was stored on a Hadoop cluster, Hadoop Cloudera distribution CDH4.6.0, Hive version 0.10.0. Regression modeling and visualization were performed with the R statistical package, version 3.2.3 (R Core Team, 2015). Graphs were produced with `ggplot2`, version 2.1.0 (Wickham, 2009). Maps were drawn with `maps`, version 3.1.0 Brownrigg (2016). Tables were created with `stargazer`, version 5.2 (Hlavac, 2015).

A.6 List and counts of broadcast news sources – Chapter 3

Publication	Scope	Attention	Protest
Associated Press	International	4851	327
CNN Wire	International	2045	232
Postmedia Breaking News	International	1351	58
The New York Times	National	944	89
St. Louis Post-Dispatch (Missouri)	Local	943	58
Daily News (New York)	Local	871	48
The Washington Post	National	760	99
UPI	International	592	23
The New York Post	Local	526	10
Gannett News Service	International	520	36
San Jose Mercury News (California)	Local	472	60
Florida Times-Union (Jacksonville)	Local	416	13
Dayton Daily News (Ohio)	Local	412	13

Continued on next page

Publication	Scope	Attention	Protest
Pittsburgh Tribune Review	Local	399	3
The Christian Science Monitor	International	369	32
USA TODAY	National	358	22
Topeka Capital-Journal (Kansas)	Local	295	10
Contra Costa Times (California)	Local	294	33
Pittsburgh Post-Gazette	Local	264	20
Palm Beach Post (Florida)	Local	263	6
Bangor Daily News (Maine)	Local	258	3
Chicago Daily Herald	Local	249	4
Deseret Morning News (Salt Lake City)	Local	225	3
The Tampa Tribune (Florida)	Local	225	6
New York Observer	Local	219	6
The Bismarck Tribune	Local	217	4
The Atlanta Journal-Constitution	Local	209	27
Spokesman Review (Spokane, WA)	Local	159	8
The Herald-Sun (Durham, N.C.)	Local	159	24
The Philadelphia Inquirer	Local	159	49
The Buffalo News (New York)	Local	150	10
Providence Journal	Local	145	8
The Capital (Annapolis, MD)	Local	136	8
Star Tribune (Minneapolis, MN)	Local	133	14
Charleston Gazette (West Virginia)	Local	127	0
St. Paul Pioneer Press (Minnesota)	Local	118	16
The Salt Lake Tribune	Local	106	4
The Philadelphia Daily News	Local	103	14
The Augusta Chronicle (Georgia)	Local	98	0
The Capital Times (Madison, Wisconsin)	Local	94	17
Lowell Sun (Massachusetts)	Local	93	0
Sarasota Herald Tribune (Florida)	Local	91	5
The State Journal- Register (Springfield, IL)	Local	89	7
Telegraph Herald (Dubuque, IA)	Local	85	0
Orange County Register (California)	Local	84	6
Intelligencer Journal/New Era (Lancaster, Penn...	Local	79	0
LNP (Lancaster, PA)	Local	70	3
Monterey County Herald (California)	Local	66	4
Vallejo Times Herald (California)	Local	64	9
Star-News (Wilmington, NC)	Local	64	4
Tribune-Review (Greensburg, PA)	Local	62	0
The Pantagraph (Bloomington, Illinois)	Local	55	0
South Bend Tribune (Indiana)	Local	54	0
San Bernardino Sun (California)	Local	53	4
Austin American-Statesman (Texas)	Local	46	5
Buffalo News (New York)	Local	45	0
Pasadena Star-News (California)	Local	45	5

Continued on next page

Publication	Scope	Attention	Protest
Sentinel & Enterprise (Fitchburg, Massachusetts)	Local	42	0
Chico Enterprise-Record (California)	Local	41	0
Whittier Daily News (California)	Local	40	3
Telegram & Gazette (Massachusetts)	Local	40	5
The Hill	National	37	0
The Berkshire Eagle (Pittsfield, Massachusetts)	Local	37	0
Daily Camera (Boulder, Colorado)	Local	33	5
Charleston Daily Mail (West Virginia)	Local	32	0
The Daily Record (Baltimore, MD)	Local	28	0
The York Dispatch (Pennsylvania)	Local	27	0
Public Opinion (Chambersburg, Pennsylvania)	Local	26	0
The Journal Record (Oklahoma City, OK)	Local	22	0
San Gabriel Valley Tribune (California)	Local	22	0
The Columbian (Vancouver, Washington)	Local	19	0
Sunday News (Lancaster, Pennsylvania)	Local	19	0
The Evening Sun (Hanover, Pennsylvania)	Local	19	0
Marin Independent Journal (California)	Local	18	3
The Daily News of Los Angeles	Local	17	0
El Paso Times (Texas)	Local	16	0
The Patriot Ledger (Quincy, MA)	Local	14	0
The Daily Record of Rochester (Rochester, NY)	Local	14	0
Messenger-Inquirer (Owensboro, Kentucky)	Local	14	0
Oroville Mercury Register (California)	Local	14	3
Eureka Times Standard (California)	Local	13	0
The Arizona Capitol Times	Local	12	0
Chapel Hill Herald (Durham, N.C.)	Local	12	0
Las Cruces Sun-News (New Mexico)	Local	11	0
Daily Tribune	Local	10	9
The Record (Stockton, California)	Local	10	0
The Bakersfield Californian	Local	10	0
The Legal Ledger (St. Paul, MN)	Local	10	0
The Bath County News-Outlook (Owingsville, Ken...	Local	10	0
Deming Headlight (New Mexico)	Local	10	0

A.7 Search strings and keywords – Chapter 3

<table>
<tr><th></th><th>Keyword or regular expression</th></tr>
<tr><td>Lexis-Nexis protest</td><td>(boycott* OR press conference OR news conference OR (protest* AND NOT protestant*) OR strik* OR rally OR ralli* OR riot* OR sit-in OR occupation OR mobiliz* OR blockage OR demonstrat* OR marchi* OR marche*) AND african american*</td></tr>
<tr><td>Twitter protest</td><td>protest(ed|ing)*|rall(y|ied|ies)|march(ed|es|ing)*|walkout|walk(ed|ing)* out|riot(ed)*|sit(-|)in|demonstrat(ed|es|ion|ing)|boycott(s|ing|ed)*</td></tr>
<tr><td>Victim keywords</td><td>ferguson, michael brown, mike brown, eric garner, freddie gray, walter scott, tamir rice, black lives matter, john crawford, tony robinson, eric harris, ezell ford, akai gurley, kajieme powell, tanisha anderson, victor white, jordan baker, jerame reid, yvette smith, philip white, phillip white, dante parker, mckenzie cochran, tyree woodson, trayvon martin, renisha mcbride, marlene pinnock, dontre hamilton, aura rosser, sandra bland, natasha mckenna</td></tr>
</table>

A.8 Creative Commons Attribution-ShareAlike 4.0 International Public License

This text can also be found at `https://creativecommons.org/licenses/by-sa/4.0/`.

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-ShareAlike 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.

c. BY-SA Compatible License means a license listed at creativecommons.org/compatiblelicenses, approved by Creative Commons as essentially the equivalent of this Public License.

d. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

e. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

f. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.

g. License Elements means the license attributes listed in the name of a Creative Commons Public License. The License Elements of this Public License are Attribution and ShareAlike.

h. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.

i. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

j. Licensor means the individual(s) or entity(ies) granting rights under this Public License.

k. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

l. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

m. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope.

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

A. reproduce and Share the Licensed Material, in whole or in part; and

B. produce, reproduce, and Share Adapted Material.

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.

3. Term. The term of this Public License is specified in Section 6(a).

4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

5. Downstream recipients.

A. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.

B. Additional offer from the Licensor – Adapted Material. Every recipient of Adapted Material from You automatically receives an offer from the Licensor to exercise the Licensed Rights in the Adapted Material under the conditions of the Adapter's License You apply.

C. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others

designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.

2. Patent and trademark rights are not licensed under this Public License.

3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material (including in modified form), You must:

A. retain the following if it is supplied by the Licensor with the Licensed Material:

i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);

ii. a copyright notice;

iii. a notice that refers to this Public License;

iv. a notice that refers to the disclaimer of warranties;

v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
C. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

b. ShareAlike.

In addition to the conditions in Section 3(a), if You Share Adapted Material You produce, the following conditions also apply.

1. The Adapter's License You apply must be a Creative Commons license with the same License Elements, this version or later, or a BY-SA Compatible License.

2. You must include the text of, or the URI or hyperlink to, the Adapter's License You apply. You may satisfy this condition in any reasonable manner based on the medium, means, and context in which You Share Adapted Material.

3. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, Adapted Material that restrict exercise of the rights granted under the Adapter's License You apply.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;
b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material, including for purposes of Section 3(b); and
c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

a. Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.

b. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or

2. upon express reinstatement by the Licensor.

For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

A.9 Typesetting

Typesetting in LaTeXwas performed with pdfTeX, Version 3.14159265-2.6-1.40.16.

BIBLIOGRAPHY

Almeida, Paul and Mark Lichbach. 2003. "To the Internet, From the Internet: Comparative Media Coverage of Transnational Protests." *Mobilization: An International Quarterly* 8(3):249–272.

Amenta, Edwin, Thomas Alan Elliot, Nicole Clorinda Shortt, Didem Türkoğlu and Burrel James Vann. 2015. "Strategies, Stories, and the Quality of News Coverage of the Civil Rights Movement in Its Heyday.". Presented at ASA Annual Meeting. Chicago, IL.

Andrews, Kenneth T. and Michael Biggs. 2006. "The Dynamics of Protest Diffusion: Movement Organizations, Social Networks, and News Media in the 1960 Sit-ins." *American Sociological Review* 71(5):752–777.

Azar, Edward E. 1980. "The conflict and peace data bank (COPDAB) project." *Journal of Conflict Resolution* 24(1):143–152.

Beissinger, Mark R. 2002. *Nationalist Mobilization and the Collapse of the Soviet State*. Cambridge University Press.

Benford, Robert D. and David A. Snow. 2000. "Framing Processes and Social Movements: An Overview and Assessment." *Annual Review of Sociology* pp. 611–639.

Benkler, Yochai. 2006. *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bennett, W Lance. 2003. "Communicating Global Activism: Strengths and Vulnerabilities." *Information, Communication & Society* pp. 143–168.

Bennett, W Lance and Alexandra Segerberg. 2012. "The logic of connective action: Digital media and the personalization of contentious politics." *Information, Communication & Society* 15(5):739–768.

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

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins and Charles Lewis Taylor. 2003. "Integrated data for events analysis (IDEA): An event typology for automated events data development." *Journal of Peace Research* 40(6):733–745.

Bond, Joe and Doug Bond. 1995. "PANDA codebook." *Center for International Affairs, Harvard University* .

Bonilla, Yarimar and Jonathan Rosa. 2015. "# Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States." *American Ethnologist* 42(1):4–17.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. "ICEWS Coded Event Data.".
URL: *http://dx.doi.org/10.7910/DVN/28075*

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2011. "Real time, time series forecasting of inter-and intra-state political conflict." *Conflict Management and Peace Science* 28(1):41–64.

Brownrigg, Ray. 2016. *maps: Draw Geographical Maps*. R package version 3.1.0.
URL: *https://CRAN.R-project.org/package=maps*

Bureau of Justice Statistics. N.d. "Data Collection: Arrest-Related Deaths.". Accessed: 2016-05-01.
URL: *http://www.bjs.gov/index.cfm?ty=dcdetail&iid=428*

Carter, Gregg Lee. 1983. "Explaining the Severity of the 1960's Black Rioting." *Unpublished Dissertation. Columbia University* .

Castells, Manuel. 2011. *The rise of the network society: The information age: Economy, society, and culture*. Vol. 1 John Wiley & Sons.

Chang, Angel X and Christopher D Manning. 2012. SUTime: A library for recognizing and normalizing time expressions.

Chatelain, Marcia and Kaavya Asoka. 2015. "Women and Black Lives Matter." *Dissent* 62(3):54–61.

Cline Center for Democracy. 2010. "The Origins of Destabilizing Events.". Accessed: 2016-04-11.
URL: *http://www.clinecenter.illinois.edu/research/publications/SPEED-Origins_of_Destabilizing_Events.pdf*

Cline Center for Democracy. 2013. "The Origins of Destabilizing Events.". Accessed: 2016-04-13.
URL: *http://www.clinecenter.illinois.edu/research/publications/SPEED-Definitions_of_Destabilizing_Events.pdf*

Conover, Michael, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer and Alessandro Flammini. 2011. Political Polarization on Twitter. In *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM)*.
URL: *http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2847/3275*

Croicu, Mihai and Nils B Weidmann. 2015. "Improving the selection of news reports for event coding using ensemble classification." *Research & Politics* 2(4).
URL: *http://rap.sagepub.com/content/2/4/2053168015615596*

Davenport, Christian. 2005. "Repression and mobilization: Insights from political science and sociology." *Repression and mobilization* .

Davenport, Christian. 2010. *Media Bias, Perspective, and State Repression: The Black Panther Party*. Cambridge University Press.

Davenport, Christian and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Davis, Gerald F, Doug McAdam, W Richard Scott and Mayer N Zald. 2005. *Social movements and organization theory*. Cambridge University Press.

Earl, Jennifer. 2003. "Tanks, tear gas, and taxes: Toward a theory of movement repression." *Sociological Theory* 21(1):44–68.

Earl, Jennifer. 2011. "Political repression: Iron fists, velvet gloves, and diffuse control." *Annual Review of Sociology* 37:261–284.

Earl, Jennifer, Andrew Martin, John D. McCarthy and Sarah Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30:65–80.

Earl, Jennifer and Katrina Kimport. 2011. *Digitally Enabled Social Change: Activism in the Internet Age*. MIT Press.

Earl, Jennifer and Sarah A. Soule. 2006. "Seeing blue: A police-centered explanation of protest policing." *Mobilization: An International Quarterly* 11(2):145–164.

Earl, Jennifer, Sarah A. Soule and John D. McCarthy. 2003. "Protest under fire? Explaining the policing of protest." *American Sociological Review* 68(4):581–606.

Eckstein, Harry, ed. 1964. *Internal War: Problems and Approaches*. Free Press of Glencoe.

Elliot, Thomas. 2014. The Cultural Consequences of Social Movements: The LGBT Movement and the Transformation of Discourse About Homosexuality. In *Paper presented at the Young Scholars in Social Movements Conference. Notre Dame, IN*.

Evans, Sara M and Harry C Boyte. 1992. *Free spaces: The sources of democratic change in America*. University of Chicago Press.

Ferree, Myra Marx, William A. Gamson, Jürgens Gerhards and Dieter Rucht. 2002. *Shaping Abortion Discourse: Democracy and the Public Sphere in Germany and the United States*. Cambridge University Press.

Finkel, Jenny Rose, Trond Grenager and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics pp. 363–370.

Francisco, Ronald. 2004. "After the massacre: Mobilization in the wake of harsh repression." *Mobilization: An International Quarterly* 9(2):107–126.

Francisco, Ronald A. 1996. "Coercion and protest: An empirical test in two democratic states." *American Journal of Political Science* pp. 1179–1204.

Franzosi, Roberto. 1987. "The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 20(1):5–16.

Franzosi, Roberto. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge University Press.

Franzosi, Roberto. 2006. *The Puzzle of Strikes: Class and State Strategies in Postwar Italy*. Cambridge University Press.

Franzosi, Roberto. 2010. *Quantitative narrative analysis*. Vol. 162 Sage.

Fraser, Nancy. 1990. "Rethinking the public sphere: A contribution to the critique of actually existing democracy." *Social text* (25/26):56–80.

Freelon, Deen Goodwin, Charlton D McIlwain and Meredith D Clark. 2016. "Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice.". Center for Media & Social Impact reported. Accessed: 2016-04-28.
URL: *http://cmsimpact.org/resource/beyond-hashtags-ferguson-blacklivesmatter-online-struggle-offline-justice/*

Gamson, William A. and Andre Modigliani. 1989. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." *American Journal of Sociology* pp. 1–37.

Garza, Alicia. N.d. "A Herstory of the #BlackLivesMatter Movement by Aliza Garza." *The Feminist Wire*. Forthcoming.
URL: *http://www.thefeministwire.com/2014/10/blacklivesmatter-2/*

Gerner, Deborah J. and Philip A. Schrodt. 1994. "Validity assessment of a machine-coded event data set for the Middle East, 1982-92." *American Journal of Political Science* 38(3):825–854.

Gerner, Deborah J, Philip A Schrodt, Omur Yilmaz and Rajaa Abu-Jabr. 2002. The creation of CAMEO (Conflict and Mediation Event Observations): An event data framework for a post cold war world. In *Presented at the American Political Science Association*. Vol. 29.

Gitlin, Todd. 1980. *The whole world is watching: Mass media in the making & unmaking of the new left*. Univ of California Press.

Goffman, Alice. 2009. "On the run: Wanted men in a Philadelphia ghetto." *American Sociological Review* 74(3):339–357.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder and Mark Woodward. 2010. "A global model for forecasting political instability." *American Journal of Political Science* 54(1):190–208.

Goodwin, Jeff and James M. Jasper. 1999. "Caught in a Winding, Snarling Vine: The Structural Bias of Political Process Theory." 14(1):27–54.

Graeff, Erhardt, Matt Stempeck and Ethan Zuckerman. 2014. "The battle for 'Trayvon Martin': Mapping a media controversy online and off-line." *First Monday* 19(2).

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* pp. 1–31.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2014. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. Technical report Working paper). Stanford, CA: Stanford University. Retrieved from http://stanford. edu/˜ jgrimmer/het. pdf.

Gurr, Ted R. 1970. *Why men rebel*. Princeton University Press.

Hamdy, Naila and Ehab H Gomaa. 2012. "Framing the Egyptian uprising in Arabic language newspapers and social media." *Journal of Communication* 62(2):195–211.

Hanna, Alex. 2013. "Computer-Aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18(4):367–388.

Hanna, Alex. 2015. "Automated Coding of Protest Events with MPEDS.". Presented at New Frontiers of Automated Content Analysis in the Social Sciences Conference. University of Zurich, July 1-3.
URL: *http://www.aca-zurich-2015.org/files/Hanna-ACA-Zurich-2015.pdf*

Harding, Sandra. 1987. Is there a feminist method? In *Feminism and methodology: Social science issues*. Indiana University Press pp. 1–14.

Hess, David and Brian Martin. 2006. "Repression, backfire, and the theory of transformative events." *Mobilization: An International Quarterly* 11(2):249–267.

Hindman, Matthew. 2008. *The myth of digital democracy*. Princeton University Press.

Hlavac, Marek. 2015. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Cambridge, USA: Harvard University. R package version 5.2.
URL: *http://CRAN.R-project.org/package=stargazer*

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

Howard, Hubbard. 1968. "Five Long Hot Summers and How They Grew." *Public Interest* 12 (Summer):3–24.

Hutter, Swen. 2014*a*. Protest Event Analysis and Its Offspring. In *Methodological Practices in Social Movement Research*, ed. Donatella Della Porta. Oxford University Press.

Hutter, Swen. 2014*b*. *Protesting Culture and Economics in Western Europe: New Cleavages in Left and Right Politics*. University of Minnesota Press.

Imig, Doug and Sidney Tarrow. 2001. "Contentious Europeans: Protest and Politics in an Integrating Europe.".

Jackson, Sarah J and Brooke Foucault Welles. 2015. "Hijacking #myNYPD: Social media dissent and networked counterpublics." *Journal of Communication* 65(6):932–952.

Jenkins, J. Craig, Charles Taylor, Marianne Abbott, Thomas V. Maher and Lindsey Peterson. N.d. "World Handbook of Political Indicators IV.". Accessed: 2016-04-06.
URL: *https://sociology.osu.edu/worldhandbook*

Jenkins, J Craig and Thomas V Maher. 2016. "What Should We Do about Source Selection in Event Data? Challenges, Progress, and Possible Solutions." *International Journal of Sociology* 46(1):42–57.

Jones, Daniel M., Stuart A. Bremer and J. David Singer. 1996. "Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns." *Conflict Management and Peace Science* 15(2):163–213.

Jungherr, Andreas. 2013. Tweets and votes, a special relationship: The 2009 federal election in germany. In *Proceedings of the 2nd workshop on Politics, elections and data*. ACM pp. 5–14.

Kadivar, Mohammad Ali and Neal Caren. 2015. "Disruptive Democratization: Contentious Events and Liberalizing Outcomes Globally, 1990–2004." *Social Forces* pp. 975–996.

Kahle, David and Hadley Wickham. 2013. "ggmap: Spatial Visualization with ggplot2." *The R Journal* 5(1):144–161.
URL: *http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf*

Kaiser, Charles. 2012. "When The New York Times Came Out of the Closet." *The New York Review of Books* .
URL: *http://www.nybooks.com/daily/2012/09/25/when-new-york-times-came-out-closet/*

Karpf, David. 2012. *The MoveOn effect: The unexpected transformation of American political advocacy*. Oxford University Press.

Khawaja, Marwan. 1994. "Resource mobilization, hardship, and popular collective action in the West Bank." *Social Forces* 73(1):191–220.

King, Gary and Will Lowe. 2003. "An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design." *International Organization* 57(03):617–642.

King, Gary and Will Lowe. 2008. "10 Million International Dyadic Events.".
URL: *http://hdl.handle.net/1902.1/FYXLAWZRIA*

Koopmans, Ruud. 2004. "Movements and media: Selection processes and evolutionary dynamics in the public sphere." *Theory and Society* 33(3-4):367–391.

Koopmans, Ruud and Paul Statham. 1999. "Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches." *Mobilization: An International Quarterly* 4(2):203–221.

Kriesi, Hanspeter. N.d. "POLCON – Political Conflict in Europe in the Shadow of the Great Recession.". Accessed: 2016-04-06.
URL: *http://www.eui.eu/Projects/POLCON/TheProject.aspx*

Kriesi, Hanspeter, Edgar Grande, Martin Dolezal, Marc Helbling, Dominic Höglinger, Swen Hutter and Bruno Wüest. 2012. *Political Conflict in Western Europe*. Cambridge University Press.

Kriesi, Hanspeter, Ruud Koopmans, Willem Duyvendak and Marco Giugni. 1992. "New social movements and political opportunities in Western Europe." *European journal of political research* 22(2):219–244.

Kriesi, Hanspeter, Ruud Koopmans, Willem Duyvendak and Marco Giugni. 1995. *New Social Movements in Western Europe: A Comparative Analysis*. University of Minnesota Press.

Kurzman, Charles and Aseem Hasnain. 2014. "When Forecasts Fail: Unpredictability in Israeli-Palestinian Interaction." *Sociological Science* 1:239–259.

Leetaru, Kalev. N.d. "Automatic Document Categorization for Highly Nuanced Topics in Massive-Scale Document Collections: The SPEED BIN Program.". Accessed: 2016-04-18.
URL: *http://www.clinecenter.illinois.edu/publications/SPEEDBIN.pdf*

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global data on events, location, and tone, 1979–2012.". Accessed: 2014-01-05.
URL: *http://eventdata.psu.edu/papers.dir/ISA.2013.GDELT.pdf*

Lichbach, Mark Irving. 1987. "Deterrence or escalation? The puzzle of aggregate studies of repression and dissent." *Journal of Conflict Resolution* 31(2):266–297.

Lohmann, Susanne. 1994. "The Dynamics of Informational Cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91." *World politics* 47(01):42–101.

Lum, Kristian and Patrick Ball. 2015. "Estimating Undocumented Homicides with Two Lists and List Dependence.". Accessed: 2016-05-01.
URL: *https://hrdag.org/wp-content/uploads/2015/07/2015-hrdag-estimating-undoc-homicides.pdf*

Maney, Gregory M. and Pamela E. Oliver. 2001. "Finding Collective Events Sources, Searches, Timing." *Sociological Methods & Research* 30(2):131–169.

Marakov, Peter, Jasmine Lorenzini, Klaus Rothenhäusler and Bruno Wüest. 2015. "Towards automated protest event analysis.". Presented at New Frontiers of Automated Content Analysis in the Social Sciences Conference. University of Zurich, July 1-3.
URL: *http://www.aca-zurich-2015.org/files/MakarovEtAl-ACA-Zurich-2015.pdf*

Martin, Brian. 2005. "The beating of Rodney King: the dynamics of backfire." *Critical Criminology* 13(3):307–326.

McAdam, Doug. 1982. *Political Process and the Development of Black Insurgency, 1930-1970*. University of Chicago Press.

McAdam, Doug. 1983. "Tactical Innovation and the Pace of Insurgency." *American Sociological Review* pp. 735–754.

McAdam, Doug, John McCarthy, Susan Olzak and Sarah A. Soule. N.d. "The Dynamics of Collective Action.". Accessed: 2016-04-06.
URL: *http://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/*

McAdam, Doug and Yang Su. 2002. "The war at home: Antiwar protests and congressional voting, 1965 to 1973." *American Sociological Review* 67(5):696–721.

McCarthy, John D., Clark McPhail and Jackie Smith. 1996. "Images of protest: Dimensions of selection bias in media coverage of Washington demonstrations, 1982 and 1991." *American Sociological Review* p. 478–499.

McCarthy, John D and Mayer N Zald. 1977. "Resource mobilization and social movements: A partial theory." *American journal of sociology* pp. 1212–1241.

McClelland, Charles. 1999. *World Event/Interaction Survey (WEIS) Project, 1966-1978, ICPSR05211-v3*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
URL: *http://doi.org/10.3886/ICPSR05211.v3*

McClelland, Charles A. and Gary D. Hoggard. 1968. Conflict patterns in the interactions among nations. In *International Politics and Foreign Policy*, ed. James N. Rosenau. University of Southern California pp. 711–724.

McCombs, Maxwell E and Donald L Shaw. 1972. "The agenda-setting function of mass media." *Public opinion quarterly* 36(2):176–187.

McKinney, Wes. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, ed. Stéfan van der Walt and Jarrod Millman. pp. 51 – 56.

McPhail, Clark and John McCarthy. 2004. "Who counts and how: estimating the size of protests." *Contexts* 3(3):12–18.

Moaddel, Mansoor. 1992. "Ideology as Episodic Discourse: The Case of the Iranian Revolution." *American Sociological Review* pp. 353–379.

Mohr, John W. and Petko Bogdanov. 2013. "Topic Models and the Cultural Sciences." *Poetics* 41:545–569.

Mueller, Carol. 1997. "International Press Coverage of East German Protest Events, 1989." *American Sociological Review* p. 820–832.

Murch, Donna. 2012. "The many meanings of watts: black power, Wattstax, and the Carceral State." *OAH Magazine of History* 26(1):37–40.

Murch, Donna. 2015. "Historicizing Ferguson: Police Violence, Domestic Warfare, and the Genesis of a National Movement Against State-Sanctioned Violence." *New Politics* XV-3. Accessed: 2016-04-26.
URL: *http://newpol.org/content/historicizing-ferguson*

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. The MIT Press.

Myers, Daniel J. 2000. "The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks." *American Journal of Sociology* 106(1):173–208.

Myers, Daniel J. and Beth Schaefer Caniglia. 2004. "All the rioting that's fit to print: Selection effects in national newspaper coverage of civil disorders, 1968-1969." *American Sociological Review* 69(4):519–543.
URL: *http://asr.sagepub.com/content/69/4/519.short*

Nardulli, Peter F., Buddy Peyton and Joseph Bajjalieh. 2015. "Climate Change and Civil Unrest The Impact of Rapid-onset Disasters." *Journal of Conflict Resolution* 59(2):310–335.

Nardulli, Peter F, Scott L Althaus and Matthew Hayes. 2015. "A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* pp. 1–36.

Newman, Nic. 2009. "The rise of social media and its impact on mainstream journalism." *Reuters Institute for the Study of Journalism* 8(2):1–5.

Oliver, Pamela. 2008. "Repression and crime control: Why social movement scholars should pay attention to mass incarceration as a form of repression." *Mobilization: An International Quarterly* 13(1):1–24.

Oliver, Pamela E. 2016. "The Ethnic Dimensions in Social Movements.". Paper to be presented at American Sociological Association annual meeting. Version: 2016-01-05.

Oliver, Pamela E and Gregory M Maney. 2000. "Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions." *American Journal of Sociology* 106(2):463–505.

Olzak, Susan. 1989. "Labor unrest, immigration, and ethnic conflict in urban America, 1880-1914." *American Journal of Sociology* 94(6):1303–1333.

Olzak, Susan. 1992. *The Dynamics of Ethnic Competition and Conflict*. Stanford University Press.

Ortiz, David G., Daniel J. Myers, Eugene N. Walls and Maria-Elena D. Diaz. 2005. "Where do we stand with newspaper data?" *Mobilization: An International Quarterly* 10(3):397–419.

O'Brien, Sean P. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12(1):87–104.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research* 12:2825–2830.

Perrin, Andrew. 2015. "Social Media Usage: 2005-2015.". Pew Research Center report. Accessed: 2016-04-29.
URL: *http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/*

Pettersson, Therése and Peter Wallensteen. 2015. "Armed conflicts, 1946–2014." *Journal of Peace Research* 52(4):536–550.

Pierce, Charles P. 2014. "The Body in the Street." *Esquire* August 22, 2014. Accessed: 2016-05-10.
URL: *http://www.esquire.com/news-politics/politics/a26327/the-body-in-the-street/*

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset Special Data Feature." *Journal of Peace Research* 47(5):651–660.
URL: *http://jpr.sagepub.com/content/47/5/651.short*

Richardson, Lewis Fry. 1960. *Statistics of Deadly Quarrels*. Boxwood Press, Pittsburgh.

Rød, Espen Geelmuyden and Nils B. Weidmann. 2014. Protesting Dictatorship: The Mass Mobilization in Autocracies Database. In *Presented at the American Political Sicence Association, Washington DC.*

Rucht, Dieter and Friedhelm Neidhardt. 1999. Methodological Issues in Collecting Protest Event Data: Units of Analysis, Sources and Sampling, Coding Problems. In *Acts of Dissent: New Developments in the Study of Protest*, ed. Dieter Rucht, Ruud Koopmans and Friedhelm Niedhardt.

Rucht, Dieter, Ruud Koopmans and Friedhelm Neidhardt. 1999. Introduction: Protest as a Subject of Empirical Research. In *Acts of Dissent: New Developments in the Study of Protest*, ed. Dieter Rucht, Ruud Koopmans and Friedhelm Neidhardt. Rowman & Littlefield.

Rucht, Dieter and Simon Teune. N.d. "Documentation and analysis of protest events in the Federal Republic of Germany, 1950-1996 (PRODAT project).". Accessed: 2016-04-06.
URL: *https://www.wzb.eu/en/research/completed-research-programs/civil-society-and-political-mobilization/projects/prodat-dokumentation-un*

Rucht, Dieter and Thomas Ohlemacher. 1992. Protest Event Data: Collection, Uses and Perspectives. In *Studying Collective Action*, ed. Mario Diani and Ron Eyerman. London: Sage pp. 76–106.

Russett, Bruce M., Karl Deutsch, Hayward R. Alker and Harold Lasswell. 1964. "World Handbook of Political and Social Indicators, 1961-1963.".
URL: *http://doi.org/10.3886/ICPSR05022.v1*

Schrodt, Philip A. 2012*a*. "CAMEO: Conflict and Mediation Event Observations Event and Actor Codebook.". Accessed: 2016-04-08.
URL: *http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf*

Schrodt, Philip A. 2012*b*. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4):546–569.

Schrodt, Philip A. 2014. The legal status of event data.
URL: *https://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/*

Schrodt, Philip A. 2015*a*. "KEDS AFP Levant Data, 1979-2015.".
URL: *http://dx.doi.org/10.7910/DVN/JUSSJX*

Schrodt, Philip A. 2015*b*. "KEDS Reuters Levant Data, 1979-2015.".
URL: *http://dx.doi.org/10.7910/DVN/7SGLJB*

Schrodt, Philip A., Erin M. Simpson and Deborah J. Gerner. 2001. Monitoring conflict using automated coding of newswire reports: a comparison of five geographical regions. In *Conference 'Identifying Wars: Systematic Conflict Research and it's Utility in Conflict Resolution and Prevention', Uppsala*. Citeseer pp. 8–9.

Schrodt, Philip A., James Yonamine and Benjamin E. Bagozzi. 2013. Data-based Computational Approaches to Forecasting Political Violence. In *Handbook of Computational Approaches to Counterterrorism*. Springer p. 129–162.

Schrodt, Philip A., John Beieler and Mohammed Idris. 2014. "Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance.". Paper presented at the ISA Annual Convention.

Schrodt, Philip A, Shannon G Davis and Judith L Weddle. 1994. "Political science: KEDS—a program for the machine coding of event data." *Social Science Computer Review* 12(4):561–587.

Scolari, Carlos A. 2012. "Media ecology: Exploring the metaphor to expand the theory." *Communication Theory* 22(2):204–225.

Sewell Jr., William H. 1980. *Work and Revolution in France: The Language of Labor from the Old Regime to 1848*. Cambridge University Press.

Sewell Jr., William H. 1996. Three Temporalities: Toward an Eventful Sociology. Ann Arbor: University of Michigan Press pp. 245–80.

Shah, Dhavan V., Alex Hanna, Erik P. Bucy, Chris Wells and Vidal Quevedo. 2015. "The Power of Television Images in a Social Media Age Linking Biobehavioral and Computational Approaches via the Second Screen." *The ANNALS of the American Academy of Political and Social Science* 659(1):225–245.

Shah, Dhavan V., Kathleen Bartzen Culver, Alex Hanna, Timothy Macafee and JungHwan Yang. 2015. "Computational approaches to online political expression: rediscovering a 'science of the social." *Handbook of Digital Politics* pp. 281–305.

Shirky, Clay. 2008. *Here comes everybody: The power of organizing without organizations*. Penguin.

Singer, J. David. 1972. "The" correlates of war" project: Interim report and rationale." *World Politics* 24(2):243–270.

Singer, Joel David and Melvin Small. 1972. *The wages of war, 1816-1965: a statistical handbook*. John Wiley & Sons.

Snow, David A., E. Burke Rochford Jr., Steven K. Worden and Robert D. Benford. 1986. "Frame Alignment Processes, Micromobilization, and Movement Participation." *American Sociological Review* pp. 464–481.

Snyder, David and William R. Kelly. 1977. "Conflict intensity, media sensitivity and the validity of newspaper data." *American Sociological Review* pp. 105–123.

Sobieraj, Sarah. 2011. *Soundbitten: The perils of media-centered political activism*. NYU Press.

Sorokin, Pitirim A. 1937. *Social and Cultural Dynamics: Fluctuation of Social Relationships, War, and Revolution*. American Book Company.

Soule, Sarah A. 2009*a*. *Contention and Corporate Social Responsibility*. Cambridge University Press.

Soule, Sarah A. 2009*b*. *Contention and corporate social responsibility*. Cambridge University Press.

Spilerman, Seymour. 1970. "The causes of racial disturbances: A comparison of alternative explanations." *American Sociological Review* 35:627–649.

Spilerman, Seymour. 1971. "The causes of racial disturbances: Tests of an explanation." *American Sociological Review* 36:427–442.

Spilerman, Seymour. 1976. "Structural characteristics of cities and the severity of racial disorders." *American Sociological Review* 41:771–793.

Strawn, Kelley D. 2010. "Protest Records, Data Validity, and the Mexican Media: Development and Assessment of a Keyword Search Protocol." *Social Movement Studies* 9(1):69–84.

Sutton, Charles and Andrew McCallum. 2006. "An introduction to conditional random fields for relational learning." *Introduction to statistical relational learning* pp. 93–128.

Tarrow, Sidney. 1989. *Democracy and Disorder: Protest and Politics in Italy, 1965-1975*. Oxford University Press.

Tarrow, Sidney. 1999. Studying Contentious Politics: From Event-ful History to Cycles of Collective Action. In *Acts of Dissent: New Developments in the Study of Protest*, ed. Dieter Rucht, Ruud Koopmans and Friedhelm Niedhardt. Rowman & Littlefield.

Taylor, Charles Lewis and David A. Jodice. 1983. "World Handbook of Political and Social Indicators III: 1948-1982.".
URL: *http://doi.org/10.3886/ICPSR07761.v2*

Taylor, Charles Lewis and Michael C. Hudson. 1972. "World Handbook of Political and Social Indicators II: Cross-National Aggregate Data, 1950-1965.".
URL: *http://doi.org/10.3886/ICPSR05027.v2*

The Dynamics of Collective Action. N.d. "Dynamics of Collective Protest in the U.S., 1960-1995. Manual for Microfilm Copying and Event Coding.". Accessed: 2016-04-14.

Tilly, Charles. 1978. *From Mobilization to Revolution*. McGraw-Hill New York.

Tilly, Charles. 1986. *The Contentious French*. Belknap Press of Harvard University Press Cambridge.

Tilly, Charles. 1995. *Popular Contention in Great Britain, 1758-1834*. Harvard University Press.

Tilly, Charles and James B. Rule. 1965. *Measuring Political Upheaval*. Vol. 19 Center of International Studies, Woodrow Wilson School of Public and International Affairs, Princeton University.

Tilly, Charles, Louise Tilly and Richard Tilly. 1975. *The Rebellious Century: 1830-1930*. Harvard University Press.

Tjong Kim Sang, Erik F and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics pp. 142–147.

Tufekci, Zeynep. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." *Proceedings of the Eigtht International AAAI Conference on Weblogs and Social Medi* pp. 505–514.

Ulfelder, Jay. 2015. ""No One Stayed to Count the Bodies"." *Dart-Throwing Chimp* . Accessed: 2016-04-18.
URL: *https://dartthrowingchimp.wordpress.com/2015/02/17/no-one-stayed-to-count-the-bodies/*

van der Walt, S., S. C. Colbert and G. Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science Engineering* 13(2):22–30.

Wallensteen, Peter and Karin Axell. 1994. "Conflict resolution and the end of the Cold War, 1989-93." *Journal of Peace Research* 31(3):333–349.

Wang, Dan J. and Sarah A. Soule. 2012. "Social movement organizational collaboration: Networks of learning and the diffusion of protest tactics, 1960–19951." *American Journal of Sociology* 117(6):1674–1722.

Weidmann, Nils B. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1):206–218.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL: *http://ggplot2.org*

Wikipedia. N.d. "List of unlawfully killed transgender people.". Accessed: 2016-05-01.
URL: *https://en.wikipedia.org/wiki/List_of_unlawfully_killed_transgender_people*

Woolley, John T. 2000. "Using media-based data in studies of politics." *American Journal of Political Science* 44(1):156–173.

Wright, Quincy. 1942. *A study of war. 2 vols.* University of Chicago Press.

Wueest, Bruno, Klaus Rothenhäusler and Swen Hutter. 2013. "Using Computational Linguistics to Enhance Protest Event Analysis.".
URL: *http://www.bruno-wueest.ch/files/Wueest_Hutter_Rothenhaeusler_2013.pdf*

Xing, Zhengzheng, Jian Pei and Eamonn Keogh. 2010. "A brief survey on sequence classification." *ACM SIGKDD Explorations Newsletter* 12(1):40–48.

Yonamine, James. 2013. A Nuanced Study of Political Conflict using the Global Datasets of Events Location and Tone (GDELT) Dataset PhD dissertation The Pennsylvania State University.

Zald, Mayer N and John D McCarthy. 1979. "Social Movement Industries: Competition and Cooperation among Movement Organizations.".