What Theories of Political Participation Can Teach Us About the Blogosphere, and Vice Versa

by

W. Abraham Gong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Political Science)
in The University of Michigan
2013

Doctoral Committee:

Professor Nancy E. Burns, Co-Chair
Professor Elisabeth Gerber, Co-Chair
Assistant Professor Eytan Adar
Professor Walter R. Mebane, Jr.
Professor Scott E. Page
Professor Nicholas A. Valentino

obama
polit
govern
presid
american
democrat
israel
senat
law
nation
comment
peopl
conserv
parti
republican
war
la
campaign
health
court
liber
elect
amp
photo
home
support
vote
washington
america
game
care
freedom
search
report
tax
public
million
feder
bill
attack
candid
district
muslim
ukip
offic
reform
immigr
movi
book
season
shop
polic
countri
council
respons
citizen
music
administr
art
travel
iraq
justic
legisl
love
congress
secur
right
summer
featur
call
continu
major
ago
isra
issu
border
hous
media
film
busi
iran
museum
onlin
site
food
econom
anti
leader
john
june
islam
tea
nuclear
http
progress
stop
protect
password
africa
truth
geert
troop
corrupt
comput
custom
review
sale
swim
unit
product
sport
eukip
student
afghanistan
burma
protest
forc
free
twitter
white
committe
gaza
label
trial
drug
organ
terror
civil
charact
world
privat
climat
activist
intern
believ
global
governor
legislatur
famili
power
seed
jpg
night
design
dog
leav
favorit
arm
poll
school
facebook
monei
billion
action
arizona
effort
middl
legal
web
paul
amend
decis
document
democraci
plan
un
award
matthew
black
watch
petit
veteran
trade
reader
defend
movement
claim
video
train
manag
magazin
program
list
beach
php
taser
viagra
post
command
crimin
clinton
research
develop
cardin
entertain
kill
point

ACKNOWLEDGEMENTS

Defending a dissertation requires great stores of inspiration, confidence, and tenacity. Fortunately, the graduate doesn't have to provide them himself.

Many thanks...

To Nancy Burns, for consistent cheerful encouragement and precisely tuned advice about methodology, participatory theory, and balancing weighty obligations.

To Liz Gerber, for many lessons about smart, compassionate teaching and mentoring, plus the finer points of program evaluation.

To Scott Page, for making math and model thinking fun, and showing me how to pitch big ideas to big audiences.

To Walter Mebane, for questioning every assumption and finding methods to go forward anyway.

To Nick Valentino, for kind and patient explanations of many mistakes and possibilities, both big and little.

To Eytan Adar, for playing the one-eyed king of computer science with good humor and good guidance.

To Mary Corcoran for persuading me to pursue a multidisciplinary PhD, and securing the support to do it right.

To Skip Lupia, Don Kinder, and Ted Brader for game theory, public opinion, and experiments.

To Rick Riolo and the assistants in the Complex Systems lab, for providing a bottomless well of computation, storage space, and debates about the merits of python.

To the National Science Foundation, Amazon Web Services, Yahoo Labs, and the Institute for Social Research at the University of Michigan, for financial support and the priceless gift of encouragement through outside validation.

Among many friends among in political science department, special thanks to Josh Gubler and Dan Magleby for being role models of patience, humor, and survival through the ups and downs of graduate school.

To Zev Berger, Dan Katz, and Charles Doriean for hours of enlightening conversations and timely distractions in the upper rooms of West Hall.

And finally...

To my mother, Susan, and mother-in-law, Rita, who took care of three wild grandkids so that I could spend summer paternity leave submerged in this manuscript. Plus all the other gifts that come with motherhood, no questions asked.

To my father, Gerrit, for being the best example of discipline, intelligence, and integrity I ever expect to know.

To my brothers, for making friendly debate a reflex, and especially to Sam, for many acts of service that made it possible to move cross-country and re-establish a family base of operations.

Last, but not least, to my wife and best friend, Erin, for her endless patience with my tendency to overthink and overcommit. Every chapter we live together leaves me loving you more.

PREFACE

What is this dissertation about?

When casual acquaintances ask the topic of my dissertation, I tell them it's about political blogging. This description is intuitive enough to paint a mental picture, and it's accurate in a literal way. My data are drawn from the blogosphere. Bloggers feature in all the research questions, models, and theories: bloggers' attitudes and demographics, blogging as a form of political participation, tone of discourse in the blogosphere, etc. Evidently—and superficially—this dissertation is "about political blogging."

However, it would be a mistake to stop there. The study of blogging also presents exciting new opportunities to answer enduring questions about politics and society. Thus, this dissertation is "about" the impact of the information revolution on politics: changes in participation, representation, and power that are unfolding as new forms of communication collide with old institutions. It is "about" norms and incentives in journalism and civic discourse, and how they affect the flow of information through society. It is "about" the scientific value of large-scale text-as-data statistical methods for understanding social systems.

In short, this dissertation is about political blogging, and several other, harder-to-explain things too. My purpose in this preface is to gather and explain these other things, in order to establish a theoretical frame of reference. This framework will be indispensible for making sense of the flood of data that will begin in the opening chapter.

With that groundwork laid, I collect a set of specific, unanswered research questions that are within methodological striking distance. An overview of these questions and my answers to them provides an outline for the remainder of the dissertation.

The practice of blogging

Blogging is the practice of maintaining a website with frequently updated entries displayed in reverse chronological order[1]. Individual entries are called "posts." In addition, most blogs allow readers to add their own comments to blog posts. In some blogs, discussion in comment threads is an important part of the medium's appeal. Several online platforms (e.g. Blogger, WordPress) allow users to publish their own blogs, free of charge. Other platforms (e.g. Google's AdWords) allow bloggers to inject advertising into their sites, making blogs a potential source of revenue for those with large enough readership.

The first blogs were created in the late 1990s by hobbyists with the web development skills necessary to create and maintain their own web sites. In 1999, Pyra labs developed the Blogger application, which streamlined and automated the process of posting updates to the web. Blogger and other automated weblogging platforms dramatically lowered the technical prerequisites to blogging, triggering a meteoric increase in the popularity of the medium (Karpf, 2008). Today, over 168 million blogs have been started (*BlogPulse.com*, 2011), and collective readership for top blogs outstrips mainstream media outlets (Smith, 2008). A 2006 Pew survey found that one in five American teens maintained a blog, and 39% of American adults read blogs (Lenhart and Fox, 2006).

[1] Since the original advent of blogging, many variations on the genre have emerged. For the sake of clarity, I follow most previous academic work (e.g. Drezner and Farrell (2008) and Pole (2009)) by defining blogs in terms of format. Hence the definition, "entries displayed in reverse chronological order."

Open questions

Most existing research on political blogging compares bloggers to journalist. In this dissertation, I follow a different thread by comparing political bloggers to political activists. If bloggers are activists, then theories of political participation can inform the study of blogging. Conversely, data from the blogosphere can open new avenues of research into political participation. Following this reasoning, this dissertation seeks to answer three broad questions about political blogging. I chose these questions because they have far-reaching theoretical implications, and are still methodologically feasible.

1. Who blogs about politics?
2. Why do they blog about politics?
3. How do bloggers' motivations affect the content of their blogs?

I will not directly try to answer questions about who reads blogs, how blogs influence mainstream media, or what effect blogging has had on specific political and policy outcomes. These are important questions, and I hope my analysis will enrich these lines of inquiry. However, as an empirical endeavor, this project focuses on bloggers themselves, not the downstream consequences of blogging. As a result, the structure of this research will most resemble studies of political communication in context, such as gatekeeping and news production (e.g. Gans, 1979), motivations behind political participation (e.g. Verba, Schlozman and Brady, 1995), and everyday political discussion (e.g. Walsh, 2004).

Validate, update, and elaborate

Answers to these questions stand to make a substantial contribution to the theory of political participation. The first question seeks to *validate* canonical theories of participation in the context of political blogging. We know a great deal about the social factors that make some people more likely to sign a petition, join a protest, or contact a government official about an issue. It seems reasonable to suppose that the same factors would lead someone to blog about the issue. However, that assertion can and should be tested.

The second question seeks to *update* theories of political participation to account for advances in communication technology. Given that the costs and skill requirements for blogging are different from previous forms of participation, it seems likely that blogging fills a different niche in the participatory repertoire. Furthermore, since access, socialization, and incentives for blogging are not distributed evenly across society, it is unlikely that the voice of the blogosphere is fully representative of the preferences of the electorate. Understanding how and why the blogosphere distorts public preferences is an important part of understanding the political consequences of this new medium.

The third question seeks to *elaborate* upon existing theories of participation. Most past studies in this area have constructed the dependent variable in terms of levels of participation: how much is one likely to get involved in politics? This is a valuable and well-developed line of inquiry. However, it ignores other normatively important aspects of participation. This limitation is driven in large part by methodological constraints on data collection, due to the field's reliance on cross-sectional surveys. Verba and Nie's (1987) older work on self-interested and community-oriented participation, Tilly's (1978) work on political repertoires, and Bowers' (2003) work on life cycle and timing in political involvement are exceptions that prove the rule. In each case, the researchers elaborated the dependent variable, and struggled with the limits of available methods and data, before arriving at interesting and suggestive findings about participation.

Specific research questions

To summarize the agument so far, I am convinced that political blogging is worth studying for two seemingly contradictory reasons. On one hand, blogging—and new media more broadly—represent an important shift in the way the body politic communicates, reasons, and decides about important political issues. These changes are worth understanding, and understanding fully, so that we can adapt our norms and institutions to the new realities of digital discourse.

On the other hand, political bloggers closely resemble activists who participate politics through other venues. To the extent the bloggers are typical of other activists, their blogs can provide insight into many aspects of political activism that are hard to observe in other venues. Thus, rich content from the blogosphere offers the chance to answer age-old questions about attention, opinion, communication, and participation in democratic society.

In short, political blogging is worth studying because it is both the same and different. My goal in this dissertation is to follow the both of these paths by investigating patterns of activism in the blogosphere and exploring their broader implications for democracy. This is the main subject of Chapter II: situating political blogging within the democratic repetoir of activism.

Hybrid methodology

The richness of data on participation in the blogosphere also provides a unique opportunity to re-imagine political participation as a phenomenon of interest. Publicly available blog archives allow us to measure levels of participation (How frequently did the blogger post, and at what length?), but they also allow us to elaborate the dependent variable by directly observing other aspects of political participation. In many ways, this is the most exciting part of the project, because it has the potential to break broad swaths of new theoretical ground.

However, in order to tap this rich new source of data, we must first develop appropriate methods. A large portion of this dissertation is dedicated to doing so. Chapter I tackles the problem of constructing representative sampling frames in dispersed online populations, using an "automated snowball census." Other studies of political blogging have relied exclusively on convenience or prominence samples that overrepresent popular bloggers and do not generalize to the population at large. Developing methods to build a sampling universe and survey a representative sample of bloggers enables us to obtain the first generalizable results about this population. It also enables us to draw direct comparisons between popular and less popular bloggers.

Chapter III introduces a data pipeline for solving a different problem: subjective content analysis at very large scales. Drawing on ideas from the literatures on traditional content analysis, combinied expert forecasts, and crowdsourced workflows, I show how we can generate reliable labels for millions of documents, even on tasks that require a high degree of subjective judgment. The chapter introduces seven codebooks for journalism and civility, and tests the process of scaling these measures using expert, novice, and automated coding processes. Although I do not take advantage of these measures in this dissertation, I strongly suspect that this kind of content analysis will make a significant contribution to scholarship in coming years. Chapter III describes enabling methods for this next wave of research.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

<u>**Table**</u>

LIST OF APPENDICES

Appendix

ABSTRACT

Few venues span the spectrum of political ideas better than the blogosphere, the sprawling online network of “web-logs” and their authors. Roughly 1.3 million Americans blog at least occasionally about politics, with aggregate daily readership exceeding that of major newspapers, and daily aggregate word counts in the tens of millions. This incredibly diverse medium captures the daily thoughts of people from all walks of life, from Senators to army wives to community activists to business owners to conspiracy theorists, all lending their voices to a public forum that was almost unimaginable a generation ago.

Previous research has focused primarily on how blogging is *different*, especially how blogging is different from traditional journalism. In contrast, I show how political blogging is strikingly similar—to political activism. The same social forces that lead people to vote, protest, or write letters to public officials can also lead them to blog about politics. Thus, bloggers are not journalists. They are activists, which means that classic theories of political participation can inform the study of blogging. This project explores these similarities, detailing the forces that drive participation in the political blogosphere, and revealing where the blogosphere represents—and distorts—the voice of the electorate. This research provides clues into behaviors that are hard to observe in other contexts, but matter deeply for society and for democracy.

Conversely, data from the blogosphere can open new avenues of research into political participation. Unlike most forms of communication, blogging leaves a permanent data trail. Archives of thousands of political blogs exist online, complete with text, dates, links, and comments. This project taps this wealth of social data using a

combination of techniques from social and computer science: survey research, content analysis, web crawling, and automated text classification. Using this interdisciplinary mix of tools, I survey hundreds of bloggers and analyze nearly eight million blog posts. In the process, I build methodological bridges between social and computer science, making software and data available for future research.

The manuscript includes three chapters.

Chapter I solves a persistent methodological problem for social scientists studying the political web: representative sampling. Virtually all existing studies of the political web are based on incomplete samples, and therefore lack generalizability. This chapter combines methods from computer science and sampling theory to conduct an *automated snowball census* of the political web. The methodology generates an all-but-complete index of English political websites. I check the robustness of this index, use it to generate descriptive statistics for the entire political web, and demonstrate that studies based on ad hoc sampling strategies are likely to be biased in important ways. In addition to creating a sampling universe for political websites, the methods and open-source software introduced in this chapter can be used to create similar sampling frames for other online content domains.

Chapter II investigates blogging as an act of political participation. In particular, we wish to understand in what ways the social predictors of political blogging resemble the predictors of participation in other modes of political activism. I find that bloggers closely resemble other activists. There are minor dfferences in some demographic characteristics, especially age and gender, and some suggestions that bloggers may follow different paths to political socialization. We also see evidence that bloggers respond to semi-professional motives in a Howewer, on the whole, bloggers fit the profile for activists remarkably well, in terms of intentions, demographics, attitudes, and behavior.

Chapter III returns to methodological issues, asking, *"How can we reliably measure* *constructive discourse in the text of political blog posts?"* This chapter describes seven codebooks used to measure important aspects of constructive discourse in blog posts. I validate these codebooks progressively with expert coders, novice coders with no prior exposure to the study, and automated coders—an exceptionally rigorous serious of tests designed to ensure that my content coding process are fully replicable, up to the scale of millions of blog posts. In the end, most (but not all) of the items from the seven codebooks pass this test, and become available for further analysis.

CHAPTER I

An automated snowball census of the political web

This paper solves a persistent methodological problem for social scientists studying the political web: representative sampling. Virtually all existing studies of the political web are based on incomplete samples, and therefore lack generalizability. In this paper, I combine methods from computer science and sampling theory to conduct an automated snowball census of the political web. This methodology generates an all-but-complete index of English political websites. I check the robustness of this index, use it to generate descriptive statistics for the entire political web, and demonstrate that studies based on ad hoc sampling strategies are likely to be biased in important ways. In future research, this bias can be eliminated by using this index as a sampling universe. In addition, the methods and open-source software presented here can be used to creating similar sampling frames for other online content domains.

1.1 Introduction

I begin with a pressing problem, captured in quotations. “The vast amount of human knowledge encoded online is the reason why the Web is such a valuable resource for politics; but ironically, the very scale of this resource makes the Web extraordinarily difficult to study” (Hindman, Tsioutsiouliklis and Johnson, 2003). “In the absence of a known population, ... a truly random sample [of relevant web sites] is not possible”

(Miller, Pole and Bateman, 2011). "Sampling the entire World Wide Web is a virtually impossible feat" (Soon and Cho, 2011).

Behind all of these quotations is an unstated understanding: without a complete index of political web sites, social scientists studying the web are denied one of their most powerful tools—sampling theory. Sampling theory is the keystone of an enormous body of social science research. It enables us to draw conclusions from manageable samples, and generalize them to whole populations. On the web—where we have lacked a valid sampling frame—we have been unable to make such claims about representative sampling and generalizability.

Instead, virtually all studies of the political web have focused on sites that are easily accessible through search engines or tracking services. Such sites tend to be prominent hubs within the online ecosystem, with frequent updates, high traffic, and relatively complex organizational structures. For some research questions, it may make sense to focus on sites where readership and attention are most concentrated. But if we are interested in using the web to study aspects of social behavior such as attention, attitudes, and speech, then ignoring the long tail of online participation is a serious oversight.

Consider the following questions, all of which are severely hampered by the lack of a representative online sampling frame.

1. Political blogging is a new form of participation in the political repetoire. Who takes advantage of this new form of participation? What distiguishes them from those who don't participate, and those who participate in other ways? Understanding the composition of the political blogosphere requires an understanding of all its participants, not just the most popular.

2. Why do some political sites become popular and others not? Answering this question is key to understanding patterns of mediated influence and selective

exposure. If we are to take this question seriously, we cannot select on the dependent variable and only investigate the sites that are already influential.

3. Most online sites receive little or no traffic. So why do their owners invest time and effort to maintain them? In many ways, the *un*popular sites present the counterintuitive behavioral puzzle.

4. How and why does the distribution of public attention over political topics change over time? Variants of this question have been the subject of speculation since the dawn of mass media (e.g. Lippmann (1927), Iyengar and Kinder (2010)). New data from the web present exciting possibilities for this line of inquiry. But a complete answer will require a full understanding of all of the participants in the political web, not just a handful of the most popular web sites.

In this paper, I show how to solve the web-sampling problem for a broad class of applications. Using a novel recombination of methods from computer science and sampling theory, I conduct an automated snowball census of the English-language political web. This technique constructs a comprehensive index of English political websites. I check the robustness of this index, use it to generate descriptive statistics for the entire political web, and demonstrate that studies based on ad hoc sampling strategies are likely to be biased in important ways.

I hope that this index of nearly 800,000 web sites will facilitate future research using surveys and content analysis to understand the Internet and its politically minded inhabitants. For future studies of online politics, representative sampling does not need to be a problem. Furthermore, any online content domain that is densely linked and can be classified accurately by its content can be indexed in similar fashion. I hope that the methods and software presented here will be useful for exploring and sampling from other domains on the web.

The paper proceeds as follows. Section two reviews past attempts to sample the political web, focusing on methodological limitations. Section three describes my methodology, in detail. Section four reports results, including web crawling statistics and robustness checks. Section five discusses the strengths and limitations of this approach. Section six summarizes and concludes.

1.2 Literature Review

The web is a potential treasure trove of data about political communication and behavior (Lazer et al., 2009). However, the scale of the available data have frustrated previous attempts at sampling (Hindman, Tsioutsiouliklis and Johnson, 2003). To the best of my knowledge, no previous study of political websites has 1) been based on a fully representative sample and 2) made use of data available on the web. All past studies have used convenience, prominence, snowball, or over-samples—strategies that cannot achieve both goals simultaneously. This section describes these common sampling strategies, gives examples, and highlights their limitations.

1.2.1 Convenience sampling

In a convenience sample, no attempt is made to ensure that the sample population is representative of the population as a whole. Instead, researchers “look under the light post” by gathering a sample that is close at hand. Examples of studies employing convenience samples include Davis’ study of Usenet discussion forums (Davis, 2009); Baum and Groeling’s analysis of news judgements on left- and right-leaning online news sites (Baum and Groeling, 2008); and Johnson and Kaye’s study of news readership among blog readers (Johnson and Kaye, 2004). In the first two studies, small sets of well-known sites were chosen for study. In the third study, an opt-in sample of blog readers was recruited for an online survey. Although studies of this kind may achieve high internal validity, they cannot draw generalizations about the population as a whole.

Some convenience samples are very large. For example, in their “meme-tracking” study of information flow, Lescovek, Backstrom, and Kleinberg mined over 90 million blogs and news media sites for permutations on quotations (Leskovec, Backstrom and Kleinberg, 2009). Similarly, Hopkins and King conducted an aggregate sentiment analysis of content from over 1 million blog posts during the 2008 presidential election (Hopkins and King, 2010). But just because a data set is large does not imply that it is representative. For example, Hopkins and King tacitly acknowledge this fact by noting that their data collection is not modeled on the representative sampling strategies used in most modern opinion polls, “just as in earlier centuries when public opinion was synonymous with visible public expressions rather than attitudes and nonattitudes expressed in survey responses.[1]” This alternative conception of public opinion is not without merit, but we should recognize that it sets aside the powerful research paradigm of representative sampling. This is a philosophical and methodological choice that should not be made lightly.

1.2.2 Prominence sampling

Prominence samples focus on the most visible sites according to some well-defined metric. Prominence samples can be used to draw conclusions about popular sites—the ones most likely to show up at the top of these rankings—but they can’t be used to make inferences about the political web in general. For example, several studies of political blogging have based their samples on lists of the most popular political blogs, according to tracking companies such as Technorati or Truth Laid Bear[2].

[1] Note that Hopkins and King are speaking of representativeness within the electorate. Within the blogosphere, their sample may be nearly “representative,” because it appears to be close to comprehensive. However, aside from noting that their data include over 1.3 million blogs, they do not make (or attempt to justify) any claims about generalizability.

[2] Sampling frames derived from search engines are convenience samples, rather than prominence samples, because methods for generating search engine rankings are customized and non-transparent. Search engine results are increasingly conditioned on user id, browsing history, geography, random usability experiments, and so on. Thus, two different users will likely see different results returned

This method is very popular. Examples of studies employing prominence samples include Adamic and Glance's network analysis of ideological clustering in popular political blogs (Adamic and Glance, 2005); Davis' work on political blogging (Davis, 2009); and McKenna and Pole's survey of "A-list" political bloggers (McKenna and Pole, 2008). Perhaps the most thorough example is Wallsten's content analysis of blog posts, which is based on a random sample of 10,732 sites featured on popular blog directories (Wallsten, 2008).

Like content analysis of newspaper articles and TV transcripts, prominence samples may be appropriate for describing trends and patterns in widely read political content. However, they are poorly suited to many other important research topics, such as selective exposure, agenda setting, and content generation. Given that one of the defining features of the web is its tremendous diversity of information, this is an important oversight. To put it another way, prominence samples focus on the handful of sites most likely to behave like traditional media outlets, rather than the vast majority of political sites that act very differently. When our research interests extend beyond A-list sites, prominence samples are inadequate.

1.2.3 Snowball sampling

Snowball samples start from a batch of known sites, then follow social ties or hyperlinks to gather other sites for the sample. Conceptually, this approach has some advantages over prominence sampling, because it potentially includes all sites, not just the most prominent. However, snowball samples have traditionally been frowned on as a kind of non-random convenience sample.

In recent years, researchers have sought to establish a sound statistical foundation for representative inference from snowball samples. Of particular note is the

from the same search query. Similarly, the same user searching for the same terms at a different time and location will likely receive different results. This lack of standardization undercuts the claim that search engine results provide a well-defined and replicable metric for determining prominence.

"respondent-based sampling" pioneered by Heckathorn. Heckathorn reasoned that small sub-populations are often connected by relatively dense social networks, making snowball-sampling a cost effective way to recruit respondents within rare or hidden populations. Furthermore, following the links in the social network in a snowball sample can be modeled as a Markov process. For networks with finite diameter, this process is ergodic, so a sufficiently long chain of referrals can ensure thorough mixing. Chain samples of this type are asymptotically independent of the initial seed sample, and can be used to derive unbiased estimates of various population statistics (Heckathorn, 1997). Alternatively, if the network and initial seed sample have certain properties (e.g. an undirected network with initial sampling probabilities proportional to edge degee), then even a short chain can produce unbiased statistics (Salganik and Heckathorn, 2004).

In general, I am sympathetic to principled approaches to snowball sampling, especially in situations where other approaches to sampling are infeasible. However, the validity of respondent-driven sampling depends on some strong assumptions, especially regarding random referrals. To quote directly from Salganik and Heckathorn:

> In order to produce analytic results about the properties of the respondent-drive sampling estimators we had to make assumptions that in some cases may be violated. For example, nonrandom recruitment of friends could influence the estimates in unknown ways. Additionally, differential recruitment success by different types of people could bias the sample of edges that we observe.

The methodology employed in this paper borrows many ideas from respondent-driven sampling. As I will discuss later, my approach uses automation and exhaustive search to conduct a full snowball census of political websites. Consequently, it shares many of the strengths of Heckathorn's methods, without having to rely on so many assumptions.

A few existing studies use snowball samples, but none that I know of has employed a Heckathorn-style model to make claims about representativeness. See for example, Hindman et. al's early work on power laws in political web sites (Hindman, Tsioutsiouliklis and Johnson, 2003); Karpf's index of "authority" among prominent blogs (Karpf, 2008); and Miller and Pole's recent content analysis of health blogs (Miller, Pole and Bateman, 2011).

1.2.4 Oversampling

The only existing studies that can make generalizable claims about online behavior are those that rely on oversampling. Oversampling starts with a very large random sample from the general population. After a short screening interview, respondents matching certain criteria are selected into the final sample. Oversampling follows a straightforward statistical methodology, but the cost of gathering a large enough starting sample is often prohibitive. In addition, any bias in answers on screener questions (e.g. from recall, social desirability, or interview fatigue) can skew the composition of the resulting sample. Also, validating survey answers against actual online behavior is very difficult—to the best of my knowledge, no study to date has attempted to do so. Consequently, this approach fails to tap the potential of the web as a rich source of political data.

Lenhart and Fox (2006) provides a good example of an oversampling strategy. Over the course of a year, they added a screener question about blogging to the end of 13,000 Pew phone interviews. The 233 respondents who said they blogged were later called back and interviewed at length about their blogging habits. This approach worked reasonably well for gathering bloggers in general, but to identify a statistically significant sample of *political* bloggers would have required a much larger sample—almost certainly too large for most research budgets.

Similar recent studies include Schlozman, Verba, and Brady's study of online participatory inequality (Schlozman, Verba and Brady, 2010), and Lawrence, Sides, and Farrell's survey analysis of blog readership (Lawrence, Sides and Farrell, 2010). Both studies rely on sub-sample analysis of large phone surveys.

1.2.5 Summary

In summary, past studies of the political web have either made use of online data, or employed representative sampling methods, but never both. This tradeoff has left us in a methodological limbo, wanting to draw inferences from rich online data, but unable to do so without sacrificing external validity.

1.3 Research Questions

This paper seeks to remedy that problem by introducing a methodology—an automated snowball census—that satisfies the requirements of sampling theory and is feasible even in the face of the technical constraints imposed by the enormous scale of the web. This approach can help resolve the tension between exploiting easily available online data, and making externally valid inferences about online behavior. Consequently, it enables generalizable conclusions about many aspects of the political web that were previously unknown, such as:

1. How many sites exist in the English political web?

2. To what extent are previous sampling frames representative of the political web as a whole?

3. How do political sites with varying levels of popularity differ with respect to organization, design, and content?

1.4 Methodology

This section describes procedures to construct an all-but-complete index of English political websites, to serve as a sampling universe in future studies of the political web.

I should set expectations appropriately at the outset: the methods described here construct a sampling index for English-language political websites, not a population of human beings. A *web site* is defined as the content found under a publically available domain-level, hypertext transfer protocol (HTTP) uniform resource locator (URL). In other words, a web site can be identified by a standard web address starting with `http://` up to the next slash: `http://www.domain-name.com/`. Thus, `http://www.google.com` and `http://images.google.com` are different web sites, and `http://www.dailykos.com/faq.html` would be a *web page* within the site `http://www.dailykos.com`.

Web sites—as opposed to other possible sampling units, such as pages or users—are attractive for several reasons. First, as described above, web sites have a precise technical definition. Barring a complete overhaul of Internet content-sharing protocols, domain-level URLs will continue to provide stable references to web content. Second, web sites are clearly linked to publically available content. Since we seek to capitalize on online content as a rich source of data, this link is essential. Third, web sites are also usually linked to identifiable social units. To the extent that we can identify the individuals and organizations that own and produce content for websites, the sites constitute records of human social behavior. This link is essential to social scientists intent on using the web as a window into psychology, attitudes, communication, and so on. Finally, websites number in the hundreds of millions: daunting, but not totally beyond the resources of social scientists to study.

Taking all these reasons into consideration, web sites are arguably the single best unit of analysis for building theories of human behavior using online data. Constructing an index of sites provides a necessary link between the online space of web content, and the social space of people and organizations. In section six, I will discuss additional steps for using an index of political web sites to draw samples based on other units of analysis.

1.4.1 An automated snowball census

First, we must construct the index. The intuition behind my approach is straightforward: all publicly available websites can be accessed over the Internet, and virtually all websites are connected by links among webpages. In principle, we should be able to build a sampling frame for the political web by following links among web sites, and classifying them one by one.

In theory, this approach works. The practical problem is time. Experts estimate that 255 million web sites existed as of 2010 (royal.pingdom.com, 2011). To borrow a phrase from Matthew Hindman, exploring so many sites would be the work "of many lifetimes." To illustrate, roughly 59,000 new websites are started every day. Some percentage of those are political. Just keeping pace with these new additions would require reading 40 new sites every minute, forever. With so many sites to search through, constructing a sampling frame by hand is not feasible.

Fortunately, this process can be automated. Instead of searching through millions of websites in person, we can write software to search through millions of websites for us. This process combines two common tools from computer science: web spiders and text classifiers.

A *web spider* is a program that explores and downloads online content by following links on web pages. Thus, the spider simulates the process of surfing the Internet by clicking one link after another. Spiders can do this tirelessly, and—when properly designed—very fast. With a good Internet connection and parallel threaded architecture, a spider can easily "crawl" dozens or hundreds of web pages in a second.

A *text classifier* is a text-as-data algorithm that categorizes documents based on the words that appear in them. Text classifiers have been used for information retrieval and natural language processing (NLP) for many years (Maron, 1961), but their use has dramatically accelerated with the rise of the Internet. Among classification algorithms, *bag of words* classifiers are the most widespread. This family of classifiers considers only the frequency or proportion of words within documents, and ignores word ordering. Essentially, such classifiers count the occurrences of common words with strong associations (or disassociations) to different categories. By aggregating scores over all the words in the document, the classifier can arrive at a very good guess as to the true classification of the document as a whole. As I will explain shortly, text classifiers can be adopted to the needs of social scientists, yielding high-reliability content coding on a virtually unlimited scale. See Manning et al. (2008) for a thorough introduction to this field. See Laver, Benoit and Garry (2003) and Hopkins and King (2010) for examples of applications in political science.

Combining web spiders and text classifiers allows us to conduct an "automated snowball census" of the political web, as follows:

1. Start with a seed batch of likely political sites[3].
2. Download this batch, and classify each site as political, or not.
3. For each political site, harvest all the outbound hyperlinks.
4. Place every previously unvisited hyperlink in the next batch of sites to be visited.
5. Repeat from step two until no new political sites are found.

[3] It turns out that the choice of seed sets does not typically matter very much. Most online networks include a connected core of sites that can all be reached from each other. As long as at least one site in that core is reachable from the initial seed site, all of the others—and all sites reachable from the core—will be as well. This result is related to Heckathorn's argument that thoroughly mixed chain samples in ergodic graphs are independent of the starting sample.

This logic behind this approach is reminiscent of Heckathorn's rationale for snowball sampling: the best place to look for political sites is close to other political sites in the network. Computational tools allow us to follow this logic to its conclusion. Every single site linked from at least one known political site is checked for political content. Since this search is exhaustive, I call it a snowball *census*, rather than a snowball *sample*.

By cataloging the sites visited in this search, we can create an index of political websites. As long as two key provisions are satisified, virtually every political website will be included in an index created this way. First, the classifier must classify accurately. Any false positives or false negatives could distort the index. Second, political websites must be adequately connected. If the political web is fragmented into disjoint islands of content, the snowball might blanket one island but never reach the others.

Fortunately, good evidence supports these provisions for the political web. As we will see in the next few sections, the process of training the text classifier and spidering the political web reveals a great deal about the reliability of the classifier and connectedness of the network along the way.

1.4.2 Hand coding

My approach to training a text classifier combines best practice from traditional content analysis with recent innovations in natural language processing. The main idea is to define political content so that it can be reliably categorized by human readers, then use statistical techniques to train an algorithm to mimic human coding. I describe these steps here in turn[4].

[4] Note: in addition to the political classifier, I also trained a language classifier to recognize English text. The process is essentially the same, and the English-language classifier is extremely accurate (100% accuracy over 200 documents), so for simplicity in exposition I focus on the more difficult task of political classification.
Perhaps surprisingly, it is easier to train a classifier to recognize languages than to recognize topics. The most common method (and the one I used here) is to classify text based on character uni-,

I define political content as content "focus[ing] on who controls power in government, and/or how that power is used." This definition works well in practice, lining up nicely with common intuition about political content, while also aligning with important theoretical concepts about the authority of the state (Burns, Schlozman and Verba, 2001). This definition excludes non-state power relationships, such as gender dynamics in the workforce[5].

Using the crowd-sourcing service Amazon Mechanical Turk, I recruited English-speaking U.S. residents to code 1,000 potentially political websites[6]. For each site, coders were given instructions, a link to the website, and a simple form to fill out[7]. Appendix A contains the instructions and codesheet. Coders were paid three cents per site, regardless of whether it was political or not. An additional 200 sites were coded four times each, in order to check inter-coder reliability. Final reliability scores were respectably high, with Krippendorff's alpha of .688 on a four-point ordinal scale and .606 on a binary scale. When coders disagreed, I used the median value as the final code for the site in the testing data. These 1,200 labeled sites served as training (n=1,000) and testing (n=200) data for the classifier.

1.4.3 Training a text classifier

I used these hand-coded sites to train a text classifier. The training procedure was designed to yield high accuracy, as well as unbiased document-level classification

bi- and tri-grams—strings of up to three characters in a row (Schmitt, 1991). For languages using non-Latin character sets (e.g. Japanese, Russian), individual letters (unigrams) can inform us that English is not the language being used. For languages using the Latin alphabet, short sequences of letters are informative of the language being used. For instance, the trigram '_de' is much more common in Spanish than English. Since a text of any length includes many trigrams, and most languages differ strongly by proportions of letter sequences used, language classifiers can be very accurate.

[5] Of course, other researchers might prefer to use different definitions. If so, this method could be replicated, starting from a different definition.

[6] These sites were a random sample of sites linked from a list of popular political web sites.

[7] Using novice coders is in keeping with best practice in content analysis (Krippendorff, 2004). Repeated iterations of coding, discussion, and correction would probably yield higher reliability statistics, but at the cost of relying on a non-transparent and difficult-to-replicate training process.

probabilities. Most existing approaches to text classification focus on the first goal and ignore the second. For this application, we want both. High accuracy gives us confidence that the classifier correctly distiguishes between political and non-political sites. Unbiased classification allows us to set thresholds and correctly weight results in later analysis.

I approach this problem using a novel, two-step procedure. Following common practice in applied text classification, we can treat documents as points in a high-dimensional feature space. In the first step, we orient a hyperplane in this space to achieve maximum classification accuracy. In the second step, we produce unbiased classification estimates by projecting and rescaling documents along this axis. This two-step procedure is useful because the data requirements are quite different for orienting the hyperplane, and rescaling document-level estimates. Similar procedures have been suggested by Hopkins and King (2010), for multi-category text classification, and Hausman, Abrevaya and Scott-Morton (1998), for regression analysis in the presence of mislabeled data. See chapter 14 of Manning et al. (2008) for further discussion and examples of the graphical intuition for hyperplane classifiers.

To formalize this intuition, it will be useful to introduce some notation. Let $i \in 1, 2, ..., n$ index a collection of documents D, and $j \in 1, 2, ..., k$ index some features, F, of those documents. Features are typically derived from words, but could potentially include other information from document text or metadata. Accordingly, each document can be described using a feature vector x_i with k entries. Each document belongs to one of two classes, denoted $y_i \in \{0, 1\}$, with $y_i = 1$ when i is political, and $y_i = 0$ otherwise. A classifier is a statistical model that defines classification probability conditional on observed features: $pr(y_i = 1|x_i)$.

I classify text using a *logistic hyperplane classifier*, applied to a feature space of 5,646 maximally informative, case-insensitive, porter-stemmed[8] words from the

[8] The Porter stemming algorithm (Porter, 2006) removes suffixes so that similar words (e.g. “Senate,” “Senator,” “senatorial”) are grouped together.

training set. From a statistical standpoint, this classifier is essentially the same as logistic regression models that are familiar in social science. Within the model, the parameter space consists of a vector of k feature weights, β, and an intercept term, α. The classifier is "trained" by estimating α and β from hand-coded documents.

$$pr(y_i = 1) = 1/\left(1 + exp(-\alpha - \beta \cdot x_i)\right) \tag{1.1}$$

As mentioned previously, I train the classifier in two steps. Step one maximizes classification accuracy by orienting the classifier hyperplane in feature space. The difficulty here is that the matrix of training data is wide and sparse. The number of word stems ($k = 5,646$) exceeds the number of documents in the training set ($n = 1,000$), and many of the words are used very infrequently. Confronted with this data structure, standard regression and logit estimators are often unbounded, and therefore fail to converge. To solve this problem, I first estimate a regularized logistic regression (RLR) model over the training data. RLR is similar to standard logistic regression, except that it includes a regularization parameter, λ, which ensures that the solution is identified, even on wide, sparse data (Ng, 2004). Essentially, that parameter penalizes beta estimates based on smaller samples, by weighting them towards zero. Equations 1.2 and 1.3 give the formal optimization problems for unregularized and regularized logistic regression, respectively[9].

$$argmax_{\alpha,\beta} \sum_{i=1}^{n} log\left(prob(y_i = 1|x_i, \alpha, \beta)\right) \tag{1.2}$$

$$argmax_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} log\left(prob(y_i = 1|x_i, \hat{\alpha}, \hat{\beta})\right) - \lambda \sum_{j=1}^{k} \hat{\beta}_j^{\,2} \tag{1.3}$$

Logistic regression has been shown to perform as well or better than other state-of-

[9] Readers familiar with RLR will recognize this expression as L_2 regularization. Other regularization terms are also possible. See Ng (2004) for a good description of the different forms of regularization.

the-art class text classification algorithms on a variety of classification tasks. With a small regularization constant, RLR asymptotically approaches the maximum likelihood linear hyperplane classifier as sample size increases. In other words, we should expect preliminary estimates of β to be close to optimal. Additionally, fast algorithms for training regularized logit classifiers have been implemented in publicly available software packages. I used the python library scikits.learn, which wraps the LIBLINEAR library for fast classifier optimization. See Zhang and Oles (2001) for a detailed technical exposition of these issues.

Table 1.1 reports estimated weights for a selection of word stems from the classifier. Observe that words like "Obama," "political," and "Senate" are strongly associated with political content, and therefore receive large positive scores. The obvious political character of weights lend prima facie validity to the training process. Words like "photo," "home," and "game" are weakly disassociated with political content, and therefore receive negative scores. Note that most documents include at least a few hundred words, and classification is performed by combining all of the words in a document. Consequently, a few misleading words in a given document are unlikely to lead to misclassification.

Despite these virtues, we cannot use the preliminary model as our final estimator for two reasons. First, RLR estimates of $prob(y_i = 1)$ are sensitive to the choice of the regularization constant, λ, which must be chosen without much theoretical guidance. I experimented with constants on the range $[10^{-10}, 100]$ and finally chose 10^{-5} because it seemed to give a small boost to classifier accuracy. However, estimated values of $prob(y_i = 1)$ were clustered unreasonably tightly around $p = .5$, with a standard deviation of only .058! Clearly, these estimates are too conservative.

Second and more importantly, the correct value of α is difficult to know until the crawl is complete. If we follow the statistical logic of case-control research design (as in chapter 5 of Agresti (2007)), α should depend on the proportion of political

Table 1.1: Estimated beta values for selected word stems

Word stem	β
obama	96.80
polit	81.72
govern	67.74
presid	55.28
american	54.96
senat	54.68
hurt	2.02
nonsens	0.94
dh	0.65
eagl	0.51
cheat	0.43
monasteri	0.37
sentinel	-0.19
wear	-0.66
kit	-2.35
filmmak	-4.77
box	-5.97
ago	-27.97
photo	-32.3
art	-32.48
home	-34.49
game	-35.1
amp	-60.4

documents in the crawled set, which differs from the proportion in the training set. Furthermore, this quantity depends on the topology of the network and the classifier itself. Consequently, α—and $prob(y_i = 1)$ for each document—can't be known prior to the crawl.

I therefore conduct the crawl based on results from the preliminary model, then recalibrate estimates after the crawl is complete, as follows. Using weights from the preliminary, RLR model, I assign each document a score, $\hat{z}_i$.

$$\hat{z}_i = \hat{\alpha} + \hat{\beta} \cdot x_i \tag{1.4}$$

This procedure projects each document onto the axis normal to the classification hyperplane. I then select a cutoff point, $\bar{z}$, and execute the snowball census, retaining any document with $z_i \geq \bar{z}$. The choice of $\bar{z}$ adjusts the search radius of the snowball census. If $\bar{z}$ is high, then only outlinks from sites that are almost certainly political will be followed, resulting is a relatively narrow search. If $\bar{z}$ is low, then even outlinks from sites that are only marginally likely to be political will be followed, resulting in a broader search. In the language of information retrieval, high $\bar{z}$ favors precision, and low $\bar{z}$ favors recall. For the purpose of constructing a sampling frame, recall is probably more important. For that reason, and to remain close to the spirit of snowball sampling, $\bar{z}$ should be selected so that $pr(y_i = 1)_{prelim}$ is not much lower than the final estimated value, $pr(y_i = 1)$, as measured in the next step. This may require educated calibration over multiple crawls. For the crawl described in this paper, I chose $\bar{z}$ corresponding to a final probability cutoff of about 40 percent[10].

Once the crawl is complete, we can conduct the second estimation step and rescale. This is done by hand-coding a small set of documents selected at random from the index. Applying unregularized, standard logistic regression values of $\hat{z}_i$ for these

[10] To make this calculation, I used Hindman's ballpark estimate for the proportion of political sites on the web (1 in 300) in the case-control correction for α given in Agresti (2007).

documents yields two adjustment parameters, $\tilde{\alpha}$ and $\tilde{\beta}$.

$$prob(y_i = 1) = 1/\left(1 + exp(-\tilde{\alpha} - \tilde{\beta}\hat{z}_i)\right) \tag{1.5}$$

These parameters can be used to construct unbiased estimates of α and β. Note that these estimates are unbiased only if we believe that the logistic hyperplane accurately describes the true data-generating process. This assumption is probably not strictly true, but given the success of hyperplane classifiers, it seems reasonable to suppose that this model is a good approximation of the underlying DGP.

$$prob(y_i = 1) = 1/\left[1 + exp\left(-(\tilde{\alpha} + \tilde{\beta}\hat{\alpha}) - (\tilde{\beta}\hat{\beta}) \cdot x_i\right)\right] \tag{1.6}$$

Applying this correction to the entire index gives us final classification probabilities. Where a binary classification is needed, sites with $pr(y_i = 1) \geq \frac{1}{2}$ are classified as political, and those with $pr(y_i = 1) < \frac{1}{2}$ are classified as non-political.

1.4.4 Classifier evaluation

Having described the process for generating an automated snowball census, we can now turn to evaluation. A classifier that makes too many mistakes cannot be trusted to categorize the whole political web. No classifier is perfect, and mine is no exception. When evaluated against human coders, my classifier for political content agrees 80.1% of the time, coding 34% of documents as political[11]. A naive reading of these results is that the classifier is "about 80% accurate." However, this puts the classifier in an unfairly negative light. When compared to each other, the same human coders agree only 80% of the time[12]. The computer agreed with the human coders more than the

[11] Following best practice in NLP, all accuracy and reliability statistics were generated by applying the classifier to a testing set of documents separate from the training set. This approach reduces the risk of overfitting.

[12] These results are from reliability tests performed on a random sample of sites encountered in the snowball census. Therefore, they best represent sites in or near the political region of the World Wide Web. For the web as a whole, the reliability of both humans and computers would probably improve.

human coders agreed with each other!

The reason lies in the definition of “political” content. When applied to the messiness of the real world, even the most clear and crisp definition has some gray area where reasonable coders can disagree. This gray area accounts for most of the difference between human coders. The computer does slightly better than the human coders because the process of training leads it to balance across the coding styles of many humans. In other words, automated coding for political content is dramatically faster than human coding, and just as accurate.

1.4.5 Homophily

The second potential problem is isolated sites. If a given site or set of sites was disconnected from the rest of the political web, the snowball spider would never find it. My first defense against this claim comes from network theory. Like many social networks, the political web exhibits a high degree of homophily. Political sites are much more likely to link to political sites. Roughly 1 in 3 of the sites linked from a political site are also political, despite the fact that only about 1 in 300 websites is political. Links among websites are not random; the best place to look for links *to* political sites is *from* political sites.

A second defense comes from the implicit scope of many web-related research projects: for all intents and purposes, isolated sites are not part of the public sphere. No other political site links to them[13]. If they have no in-links, then no search engine can index them. Without search engine traffic or links, the only people who could visit such sites are those who already know their exact web addresses. Therefore, it seems fair to say that posting to an isolated site is a private action, not an act of public political participation.

[13] The crawler used in this paper explores only the index page of any given web domain. Consequently, it is possible that there exist sites that are not linked from any index page, but are linked from internal pages. Results shown in the next section suggest that sites of this type are few and far between, and mostly inactive.

1.4.6 Summary

So far, I have described a process for conducting an automated snowball census of the political web. Conceptually, the same approach should work for any subdomain on the web, as long as (1) the text classifier is accurate, and (2) the network is sufficiently connected. The political web seems to meet both of these criteria. In the next section, I describe results from the crawling process and supply additional robustness checks.

1.5 Results

I implemented the process described above in python. SnowCrawl, an open-source python module, provides a common API for directed webcrawls using a single process, multiple processing, or a client-server architecture. SnowCrawl also automates storage of downloaded files, edge lists, and state backup. Source code, examples, and documentation are available in a google code repository: `http://code.google.com/p/snowcrawl/`.

For a snowball census conducted in multiprocessing mode in August of 2010, the code executed in less than 24 hours, crawling some 1.8 million sites, and classifying about 800,000 as political.

This census was intended for evaluation purposes. Consequently, I started from a relatively small seed set of 311 sites selected at random from the Technorati index of top blogs. Using a small seed set allows for strong tests of the snowball census methodology. We deliberately set aside existing lists of political sites, so that they can serve as comparison samples later on.

1.5.1 Robustness checks

Does it actually work? Without another census to compare against, comprehensive tests are impossible. However, we can run some “back-of-envelope” validity checks.

First, we can ask if the web spider found about the right number of political sites. Older work (Hindman, Tsioutsiouliklis and Johnson, 2003) based on patterns of browsing traffic on the Internet, placed the percentage of political sites around a third of a percent. Given last year's estimate (royal.pingdom.com, 2011) of 255 million web sites and monthly growth of 7.1 million sites, we should expect to find about 826,000 political sites in our August crawl. The total count from my snowball census is in just the right ballpark: 789,818 political sites.

As a second robustness check, we can look to see if any known political sites are obviously missing or misclassified. Prominent political sites show up early in the sample: The Huffington Post, Daily Kos, whitehouse.gov, and so on. A quick search within the index also reveals 497 house.gov sites and 217 senate.gov sites in the census. These are the official websites of U.S. Congressmen, Senators, and committees—it appears that the classifier found all of them.

Taken together, these checks provide reasonably strong evidence that this index is close to complete. Next, I compare this sampling universe to those used in previous studies.

1.5.2 Comparison with previous methods

In order to makes these comparisons, I obtained the sampling frames used in several of the largest existing studies of the political web: the entire listings of political blogs from the Yahoo! Directory, blogcatalog.com, and technorati, as well as the blog URLs used in Adamic and Glance's study of political homophily in the blogosphere (Adamic and Glance, 2005). Although the sampling units are slightly different (political sites versus political blogs), these lists of political blogs constitute the best available baseline for comparison to the automated snowball census.

How does the census compare? Judged in terms of volume, the census is certainly a more complete index. The four blog lists include 385, 4469, 2700, and 1490 blogs,

Table 1.2: Comparison to previous sampling frames

index	n	Included	Broken	Ineligible	Inactive	Recall
Yahoo Directory	385	70.39%	5.71	5.71	2.86	89.14
Blogcatalog	4469	49.97	11.59	21.14	2.27	76.11
Technorati	2700	73.85	3.37	15.36	0.75	92.49
Adamic and Glance	1490	48.39	26.05	8.14	9.07	87.03

respectively. In contrast, the census includes some 470,000 blogs[14], far more than any previous study of blogs.

However, it might be the case that the census simply finds different political blogs, and that none of the sampling universes is complete. To guard against this possiblity, I searched the index generated by the automated snowball census to see what proportion of sites was included. Initial results are strong, but not entirely encouraging. As many as half of the URLs in the smaller lists were not included in the census. The first two columns of Table 1.2 report list sizes and results from this search.

In order to make sense of these discrepancies, random samples were drawn from each list. The URL and website of each sampled site *not* found in the snowball census were inspected manually. As it happens, most of the sites not included in the census were missing for one of the following reasons: the URL pointed to a broken, nonexistent, or non-public (e.g. requiring password for access) site; the site was not English or contained no political content; or the site was inactive at the time of the crawl. All of these considerations place the blog URLs outside the intended scope of the crawl. When we remove broken, ineligible and inactive URLs from consideration, the final results are much more favorable: the census included between 76 and 92 percent of the political blogs in each list[15]. This level of coverage is comparable to the percent of U.S. households reachable by phone using a landline (Blumberg, 2011). Conversely, none of the other lists contains more than half a percent of the sites found

[14] This estimate comes from the sample and content analysis described in the next section.

[15] The list with the worst coverage here is blogcatalog, which allows users to self-subscribe: bloggers can add their blogs to this list even if they have no other in-links. Consequently, a blog might be featured on this list even if it fails the "public sphere" criterion mentioned earlier.

in the snowball census. Although not perfect, the snowball census clearly provides coverage is far superior to any online sampling frame existing previously. No available online sampling frame is perfectly comprehensive, but the automated snowball census is far more comprehensive than any previous list.

As an additional comparison, we can consider the distribution of political web sites by popularity within each sample[16]. Figure 1.1 shows kernel density estimates for each of these distributions, measured by logged in-degree. For convenience in visualization, the maximum height of each distribution has been normalized to unity.

It is readily apparent that these lists are all skewed towards sites with far more more in-links than average. The absolute differences are quite large. On average, a political site from the snowball census has 89 in-links. In the other lists, the average value ranges from 304 (BlogCatalog) to 796 (Yahoo)[17]. In terms of sampling probabilities, the Yahoo index is 54 times more likely to include a site if it is among the top 5,000, and 85 times more likely if the site is in the top 500. Previous studies based on these lists have dramatically under-represented less linked-to sites.

All of this evidence strongly supports the conclusion that the new index is a superior sampling universe for political websites. It includes vastly more political sites, including large majorities of the sites on the best available comparison lists. Moreover, when compared to the snowball census, it is clear that previous methods have dramatically over-represented the most popular sites, largely ignoring the long tail of the distribution.

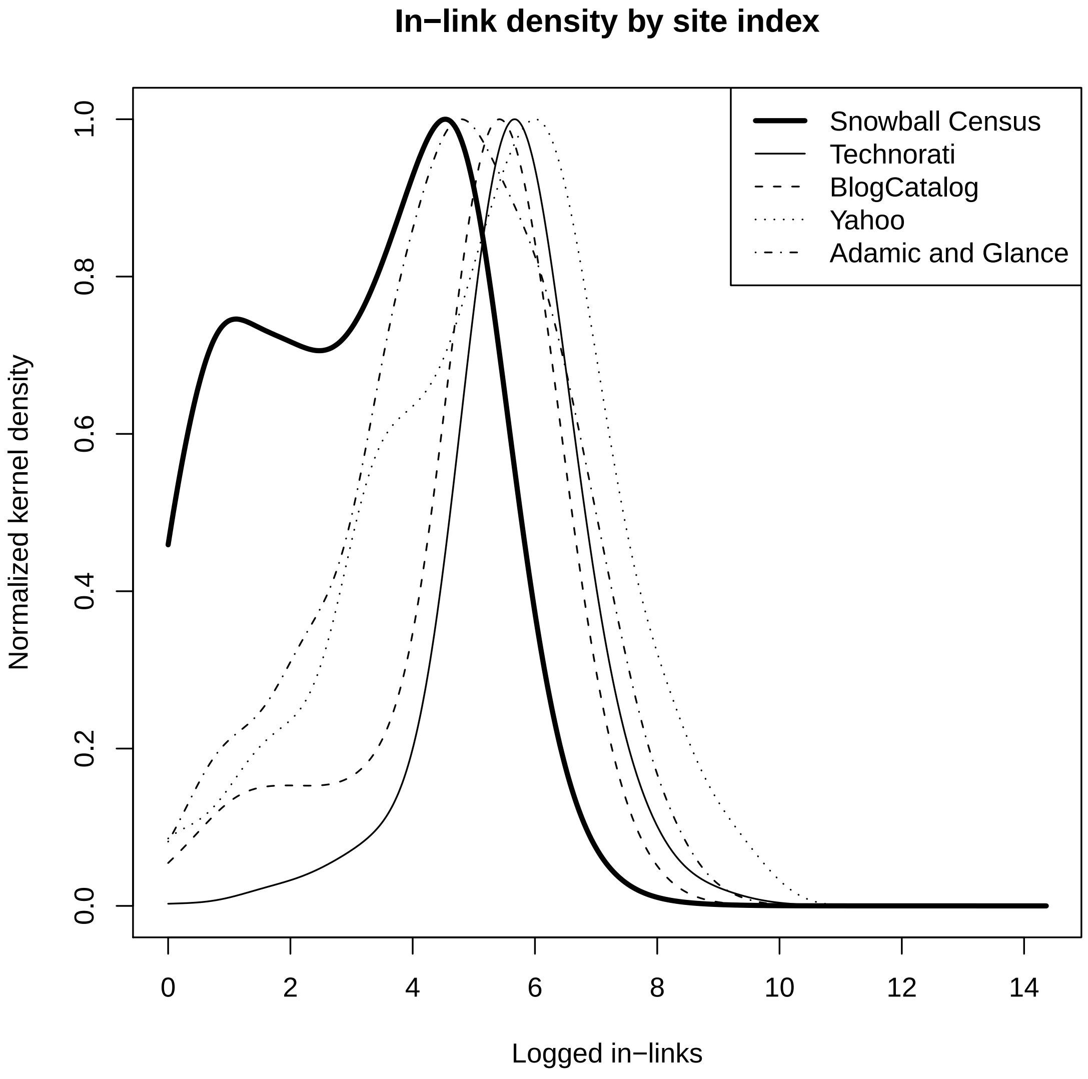
1.5.3 Descriptive statistics

With this index in hand, we can generate the first fully representative description of the political web. To do so, I sampled 150 websites from each of three strata

[16] Hindman has shown that in-degree, search engine ranking, and web traffic are all closely correlated (Hindman, Tsioutsiouliklis and Johnson, 2003). I rely on in-degree within the snowball census as a measure of site popularity.

[17] These values include self-loops–links from sites to themselves, and are therefore somewhat inflated.

Figure 1.1: In-degree density for political blogs in different sampling frames



within the census: the top 500, top 5,000, and full census of political websites. Strata were determined by inlinks from sites found in the crawl[18]. For each site, workers on Mechanical Turk were paid $0.10 to fill out a short codesheet including questions about the ownership, organization, and content of the site.

Table 1.3 reports descriptive statistics for each stratum. For the purposes of describing the political web, the important data are in the rightmost column, which reports percentages for a sample of the full census. Thus we see that 55.6 percent of political websites are personal websites run by individuals or informal groups, as opposed to websites run by organizations such as news media outlets, political campaigns, corporations, etc. 59.5 percent of political sites are formatted as blogs; 62.2 percent have more than one author; only 6.1 percent are updated multiple times per day.

I also asked about certain design elements within pages. Half of sampled sites include advertising, and nearly half include a “blogroll” or collection of links to related sites. About one in five political sites features video content. Forty percent include identifying information about their authors. Somewhat surprisingly, nearly a third of sites include buttons or forms soliciting donations or recruiting volunteers.

In terms of political content, polls, public opinion, and elections appear to be the most popular topics, followed by legislation and law-making, implementation and execution of public policy, philosophical discussion of the role of government, state and local government, political figures, and political parties. In general, foreign policy, international relations, and decisions by courts receive less attention on political sites.

1.5.4 Comparison across strata

Since all previous studies have used less than fully representative samples, it is reasonable to ask what difference, if any, a more complete sampling frame would have

[18] As Hindman (2010) has shown, inlinks are strongly correlated with other measures of popularity, such as traffic and search engine rankings.

Table 1.3: Characteristics of political websites by strata

	Top 500	**Top 5,000**	**Census**
		Organization	
Personal websites	33.9%***	46.9%	55.6%
Owned by organizations	66.1***	53.1	44.4
Formatted as blogs	51.4	70.5	59.5
Multiple authors	75.2*	66.7	62.2
Multiple updates per day	43.4***	19.4***	6.1
Updated less than weekly	14.2***	21.4***	42.7
		Design	
Advertising	67.3**	57.1	51.2
Blogroll	57.5*	66.3***	45.1
Videos	48.7***	35.7***	18.3
Identifying information	47.8	50.0	41.5
Forms for donations, etc.	36.3	32.7	30.5
		Content	
Polls and public opinion	70.8***	65.3*	52.4
Elections and campaigns	50.4	45.9	51.2
Legislation and law-making	43.4	41.8	43.9
Implementation of policy	38.1	39.8	30.5
Decisions by courts	34.5***	24.5	17.1
Political figures	46.0***	39.8**	24.4
Political parties	38.9***	32.7*	20.7
Philosophical discussion	26.5	29.6	25.6
State and local government	36.3*	38.8**	24.4
Foreign policy	42.5***	38.8***	15.9
International relations	31.9**	33.7**	18.3

Cell entries show the percent of sites that have certain organizational characteristics, or contain design element or content types. Stars indicate significance levels in pairwise t-tests between Census results and the Top 500 or Top 5,000 strata $(*p < .1)(**p < .05)(***p < .01)$. Many aspects of political websites differ significantly between sites in the head and tail of the distribution.

made. Since previous studies overrepresent popular sites, I compared popular sites in the head of the distribution against the entire population of English-language political websites.

The first and second columns in Table 1.3 allow us to make these comparisons. Each value that differs significantly from the corresponding census value (as measured by a pairwise t-test) is marked with an asterisk. Each entry marked as such is a likely source of bias in previous studies of the political web.

Thus, we see that the top 500 websites are more likely to be owned by organizations and maintained by multiple authors, with far more frequent content updates. Popular sites are also more likely to feature ads, links to other relevant sites, and video content. In this sample, they were slightly more likely to solicit donations and volunteers, but the difference between strata was not statistically significant.

A-list sites also differ in the kinds of content they cover. Popular sites include more types of content overall: 4.74 types for the top 500, and 4.50 for the top 5,000, versus 3.46 for the full census. Most of the difference comes from increased coverage of foreign policy, political figures, polls and public opinion, and court decisions. These differences in attention are substantively large: compared to sites in the full census, a top 500 site is twice as likely to discuss decisions by courts and nearly three times as likely to discuss foreign policy.

These results underscore the importance of proper sampling. Without a proper sampling frame, substantive conclusions about the political web are likely to be biased—in some cases, severely.

1.6 Discussion

This section discusses contributions, limitations, and directions for future research.

1.6.1 Contributions

Given that many aspects of snowball census are similar to snowball sampling, it is reasonable to ask what we gain by using a census. My answer is twofold. First, the validity of a census relies on fewer assumptions. Current statistical frameworks for snowball sampling rely on assumptions about random recruitment among social ties, and the asymptotic mixing properties of random walks on social networks. In contrast, a snowball census relies only on the weaker condition of connectedness within the subpopulation of interest.

Second, and perhaps more importantly, a census is easier for non-technical researchers to use. Constructing the index takes specialized software and computing power, but once generated, it can be used as needed, with little technical expertise. Like a phonebook provides an off-the-shelf method for sampling in phone surveys, this index provides an easy way to sample from the political web. The full census and various sub-samples of interest are available for download at `http://www-personal.umich.edu/~agong/resources.html`.

Furthermore, researchers with the requisite programming expertise have the option of running the snowball software again. In this case, different classifiers can be trained and used to explore different subdomains on the World Wide Web.

1.6.2 Improving coverage

The simplest way to improve coverage from the automated snowball census would be to expand the list of seed sites. For the sake of evaluation, I seeded the crawl with a small batch of political sites. The resulting index contained most, but not all of the sites in other indices. Of course, we can easily include all those other sites in the seed set for subsequent crawls, guaranteeing that they are included in the overall census.

Another simple way to improve the coverage of the census would be to conduct repeated crawls. The crawl evaluated here was conducted once, scanning the front pages of political sites within a 24-hour period. If a given site was not linked from another site's front page on that day, or didn't feature political content on its front page, it might have been missed. Executing repeated crawls would improve coverage among sites of this kind.

Both of these improvements to the methodology are relatively straightforward, requiring no new training data, software, or evaluation.

1.6.3 Classifier accuracy

As described above, the classifier described here performs quite well—as well as human coders. However, there are several means by which incremental improvements in classifier accuracy might be obtained. First, better instructions and training for human coders could eliminate some errors in the training data used to calibrate the classifier. This approach would likely improve both human-human and human-computer reliability scores. Second, in a similar vein, additional training data would likely lead to modest improvements in classifier accuracy. Third, an expanded feature space including more words or perhaps bigrams and trigrams, would probably improve the classifier a little bit as well. Fourth, experimentation with different classifiers (e.g. nonlinear SVM kernels) might also improve the classifier slightly. Overall, given the already-high accuracy of the political classifier, we should expect incremental improvements in performance at best.

In a more promising direction, the classifier could be redesigned to incorporate other forms of information. At present, the classifier only makes use of text on the main page of a web site. Consequently, short pages (i.e. those containing less than 100 words) offer less material for classification and are more likely to be misclassified. With some additional effort, the classification algorithm could be trained to incorporate text from other pages within the site, the structure of the hyperlink network surrounding the site, and so on. For sites with few words, these additional information sources might substantially improve classification accuracy.

1.6.4 Alternative units of analysis

An automated snowball census generates an index of web sites. Of course, for some applications, we might prefer other units of analysis: users, voters, blogs, web pages, etc. As argued previously, sites provide a bridge between socially meaningful and content-based units of analysis. Therefore, the index described here can help generate sampling frames for these units as well. However, the conversion requires some extra work.

For example, consider a population used in several previous studies: political blogs. As a first approximation, we might assume that blogs are a strict subset of web sites. In that case, we can obtain a random sample of political blogs by drawing a random sample of political web sites from the index, and filtering out non-blogs[19]. Since this approach allows us to sample first and filter afterwards, we only have to search through a few hundred or thousand blogs—not the whole web. As a result, using the index is much easier than starting from scratch.

A more nuanced approach to sampling political blogs would acknowledge that the mapping from sites to blogs is not one-to-one. Some sites, such as The Daily Kos and Fire Dog Lake, host multiple blogs in a "diary" format. In this case, we can employ a cluster sampling strategy, treating political sites as clustering units, and drawing subsamples of blogs from sites as needed. Weighting will be more complicated, but as long as we are confident that all (or virtually all) political blogs are contained within political web sites, this approach will generate representative samples.

A third challenge would be to sample political bloggers, rather than blogs. In this case, we could generate a sample of blogs using the process described above. Following the usual process for sampling within households, we could identify all

[19] Depending on the intended application of the sample, it might be desirable to weight on the probability of filtering (e.g. conditional on in-degree).

bloggers associated with a given blog, and use a Kish grid to select among them. Bloggers who maintain more than one blog could be reweighted appropriately.

The common thread through these examples is that starting from a comprehensive index of political web sites can simplify sampling, even when sampling on a different unit of analysis.

1.6.5 Directions for future research

At the end of a technical and methods-oriented paper, it is worth pausing to consider the value of the methods to the field at large. The web is one of the richest sources of political data in all history, capturing the ideas, opinions, messages, and reactions of a broad cross-section of society in networked panel form. Out of necessity, previous studies of the political web have focused on a small percentage of easily accessible popular web sites. A full census of political web sites, plus a methodology for constructing similar sampling frames for other online content domains, equips us to more effectively interrogate and draw social inferences from the web. Following are thoughts on future directions for research in this area.

First, rich description of the political web. With a sampling frame in hand, it should be easy to draw a sample, measure the properties of web pages and their authors using content analysis and/or surveys, then make generalizable inferences to the political web as a whole. This paper includes preliminary results in this direction, but much work remains to be done.

Second, investigation of organizational structure within sites. Results presented here hint at the diversity of organizational structures in the political web. Different sites have different ownership structures, design elements, and means of content production. Understanding this heterogeneity in online content production will be essential to making valid inferences about the attitudes, incentives, and goals of content producers.

Third, a more thorough study of network properties. The homophily and edge density statistics presented here offer only a cursory analysis of the structure of the political web. Because the snowball spider automatically generates an edge list for all sites searched, it opens exciting possibilities for studying the political web as a complex network. Such research could add greatly to our understanding of information flow, social connectivity, and political involvement.

Fourth, change and stability over time. The snowball census takes a snapshot of the political web. Revisiting political web sites across months would create a longitudinal panel uniquely suited to measuring changes in attention, attitudes, and speech in the public sphere. Given the difficulty of conducting direct experiments in this sphere, such a panel is a promising vehicle for testing causal claims.

1.7 Conclusion

In summary, this paper demonstrates how a combination of tools from computer and social science can be used to conduct an automated snowball census of the political web. I have argued that this process generates an all-but-complete index of the political web, providing strong theoretical and empirical support for that claim. Consequently, it seems reasonable to use the resulting index as a sampling universe for the political web, solving a persistent methodological problem in recent social science.

CHAPTER II

Blogging as an act of political participation

2.1 Introduction

In the last decade, blogging—maintaining a web page of posts in reverse-chronological order—has developed as an important part of the American political landscape[1]. Political bloggers number in the hundreds of thousands, with aggregate daily readership exceeding that of major news outlets (Smith, 2008). This shift in the political information environment has been linked to important changes in the dynamics of American politics: the decline of traditional media (Meyer, 2009), disintegration of traditional norms of journalism (Keen, 2007), increasing political polarization (Prior, 2007), the rise and fall of prominent politicians and journalists (Drezner and Farrell, 2008; Hindman, 2010), and changes in elite tactics for fundraising, constituent services, and mobilization (Coleman, 2005). Many debates about the structure, reach, and impact of the political blogosphere remain open.

In order to understand the blogosphere, its evolution, and its likely long-term impact on politics, we must understand the social forces that drive this new medium. This chapter contributes to this line of research by situating political blogging within

[1] Blogging, social networking, and other new media seem to be having some impact in politics elsewhere around the world as well (Rheingold, 2003) (Coleman, 2005) (Harb, 2011). For the purposes of this paper, I set aside questions about the impact of new media in authoritarian regimes, parliamentary systems, and so on. Instead, I focus squarely on the political outlier that is the United States.

the democratic repetoire of participatory acts. Within this framework, we describe the contours of political blogging as a mode of participation, compare and contrast blogging to other forms of activism, and explore dimensions of participation that are difficult to study in other contexts. This combination of representative description and direct hypothesis testing opens new avenues for understanding the relationship between activism, journalism, and new media.

To this end, I constructed the first representative sample of active, English-language political bloggers (Gong, 2011) and surveyed these bloggers alongside a comparison sample of U.S. residents. This case-control research design (Schlesselman and Stolley, 1982) enables direct comparisons between representative samples of bloggers and U.S. residents at large. Stratification within the sample enables further comparisons of bloggers by level of popularity.

In terms of substantive conclusions, I find that the decision to participate in the blogosphere is largely explained by the same forces that explain other modes of participation: engagement with politics, mobilization through social networks, and resources (e.g. time, money, and skills). However, the balance of relevant resources differs somewhat for blogging. In particular, whether a person blogs is strongly influenced by political interest and education, and only barely by income and traditional measures of civic skills. Relative to other kinds of activism, political blogging skews younger and male.

Setting aside these differences in emphasis, political blogging fits comfortably within the resource theory of political participation. Additional supporting evidence comes from the fact that bloggers' expressed motives for blogging are largely consistent with activist motives. This conclusion is further strengthened by the close resemblence that political bloggers bear to offline activists.

When we draw comparisons among bloggers by levels of readership, some unexpected patterns emerge. Not surprisingly, popular "A-list" bloggers report substantially more attention from policymakers and the mainstream media. They also put much more effort into their blogs, and earn substantially more revenue. However, there is no statistically significant relationship between popularity and ideology, partisanship, offline participation, or self-expressed reasons for blogging. These results cast light on the increasingly blurry professional boundaries around journalism and activism, and open intriguing possibilities for studying the flow of information through democratic politics.

The chapter proceeds as follows. The next section reviews the literature on political blogging, especially studies relating to blogging as an act of political participation. Section three draws lays out a set of seven research questions to investigate political blogging from several complementary directions. Section four describes my methodology, with special attention to collecting comparable survery data, measuring popularity, and sample weighting in the political blogosphere. Section five relates results for our seven research questions; section six discusses and synthesizes these results with existing theories of political participation and media. Section seven concludes.

2.2 Literature review

This section reviews the relevant literature on political blogging, with emphasis on blogging as a mode of political activism. The first half of the section reviews early hopes and fears for the new medium of blogging, and the evidence to date. The second half focuses on theories that categorize and explain political bloggers' behavior. It compares and contrasts the two primary frames (blogging as journalism, versus blogging as activism), argues that activism is the better frame, and points out limitations within existing blogging-as-activism research.

2.2.1 Early hopes and fears

When "new media" first emerged, proponents predicted utopian impacts. Imagining the great new possibilities of a World Wide Web, analysts and commentators rhapsodized about the wisdom and creative potential of crowds (Surowiecki, 2005; Howe, 2009; Tapscott and Williams, 2008), and the wealth latent in networks and long tails (Benkler, 2006; Anderson, 2008). In the new ecology of online media, "web-logs"—a new genre of web site featuring frequent posts displayed in reverse-chronological order—had a particular role to play as town criers. Blogs, it was thought, would invigorate participation, transparency, and accountability in industry and government by empowering many-to-many communication among a critical mass of citizens (Shirky, 2009; Jarvis, 2009). Bloggers themselves derided "Big Media" as technologically obsolete and institutionally defunct (Reynolds, 2007; Gillmor, 2006).

As the positive effects of blogging failed to materialize on the scale first imagined, a negative backlash developed. Some worried that citizens obsessed with the ideological convenience of "the daily me" (Negroponte, 1995) would wrap themselves in insular cocoons of likeminded opinion (Bishop, 2009). Within this framework, blogging came to be seen as a vehicle of selective exposure, partisan vitriol, and ideological Balkanization (Sunstein, 2007; Prior, 2007). Others worried about a rising tide of intellectual mediocrity, with attendant civic disengagement from difficult issues (Keen, 2007; Bauerlein, Walesh et al., 2009; Carr, 2011); or an Internet-driven centralization of control over information in society (Postman, 1993; Carr, 2008; Zittrain, 2009).

Although these debates remain unresolved, several consequences—good and bad—of blogging have now been well-documented. Perhaps the clearest change is in the offline media environment: loss of advertising revenue to online media has driven many paper-and-ink newspapers out of business (Chen 2009; Arango 2009). Bloggers are probably partly responsible (Karpf, 2008); in any case, they have filled an information niche by attracting a broad base of readers. Daily blog readership may be as high as

24 million per day (Smith, 2008), and revenue for advertising and subscriptions in the blogosphere is in the hundreds of millions of dollars (White and Winn, 2009)[2].

From this vantage point, bloggers have taken an active role in politics (Coleman, 2005). Political bloggers were largely responsible for Trent Lott's precipitous fall from power in the Senate after his purportedly racist comments in 2002, Howard Dean's unexpected emergence as a serious competitor in the 2004 Democratic primaries, and Dan Rather's loss of credibility and eventual resignation after the controversy over forged memos he presented on 60 Minutes (Davis, 2009). These examples demonstrate that bloggers have the ability to keep issues on the media agenda, raise awareness (and funds and volunteers) for causes and campaigns, and put substantial pressure on prominent politicians and journalists, at least some of the time (Drezner and Farrell, 2008).

2.2.2 Blogging as activism versus journalism

Early academic and popular work frequently asked, "Are bloggers journalists?" Of course, the conclusion depends on one's definition of journalism, but for the most part, the answer was a resounding no. Most bloggers do not see themselves as journalists (Lenhart and Fox, 2006), and the vast majority of bloggers have audiences much smaller than typical journalists[3]. Moreover, the content, practices, community, and norms of political blogging differ in clear and important ways from those of mainstream media (Perlmutter, 2008; Davis, 2009; Pole, 2009).

A second wave of research encouraged a kind of detente between bloggers and mainstream journalists. The two information channels fill different roles, both crucial to the modern news media ecology (Lasica, 2003). Following this reasoning, researchers

[2] Precise estimates of blog readership and revenue are hard to come by. Estimates for the political blogosphere are even harder. The numbers reported here are conservative projections from Pew and Technorati data for the whole blogosphere.

[3] Hindman (2010) observes that "only a few dozen blogs get as many readers as a typical college newspaper."

have asked how bloggers compare to traditional journalists in their ability to attract audience share (Graf, 2006), inform and persuade readers, frame issues (Woodly, 2008), and ultimately influence political and policy outcomes (Drezner and Farrell, 2008). Prominent topics include the credibility of blogs relative to traditional media (Johnson and Kaye, 2004) and the role of blogging in agenda setting (Wallsten, 2008; Leskovec, Backstrom and Kleinberg, 2009).

In recent years, the perspective has begun to shift. Instead of treating bloggers as journalists, some researchers have begun to study bloggers as activists[4]. Given that bloggers are engaging with political issues in a public forum, it seems reasonable to think of political blogging as a form of political participation (Karpf, 2008). According to this line of thought, blogs have joined votes, boycotts, protests, petitions, etc. in the repertoire of political behavior (Tilly, 2004).

This "blogging as activism" approach is satisfying because it grounds political blogging in existing models of activism. However, systematic empirical work on blogging as a form of activism is just beginning. To date, two approaches have dominated this field: descriptive analysis, and tests of cannonical models of participation. For reasons I will discuss, neither of these approaches provides a fully satisfactory method for explaining why bloggers do what they do.

Most of the early studies of political blogs and blogging had a descriptive, exploratory flavor. They were intended to introduce readers to the format, participants, norms, and folkways of the blogosphere. See Karpf (2008), Wallsten (2008), Pole (2009), and Davis (2009) for particularly thorough studies of this kind. This "guided tour" approach is entirely appropriate for acquainting and orienting readers to the new world of online political participation. These authors' analysis of content in the blogosphere is particularly noteworthy, since it captures an important aspect of

[4] A note on terminology: following Burns, Schlozman and Verba (2001), I will use the words "activism" and "participation" more or less interchangeably to mean "activity that has the intent or effect of influencing government action."

activism—what activists say—that is difficult to observe in other contexts, and has therefore been undertheorized. With that said, it is also worth noting several important features that most studies in this area are missing: representative sampling and the resulting ability to draw generalizable conclusions; reliability statistics and procedures for replication; and—most importantly—explicit testing of important hypotheses.

On the other hand, a second branch of research focuses on testing canonical models of political participation (Verba, Schlozman and Brady 1995; Rosenstone and Hansen 1993)in online modes of activism. If political blogging is a form of activism, then these models predict that involvement in the political blogosphere is largely a function of one's political engagement, resources (time, skills, and money), and recruitment through campaigns and social networks.

To date, the most thorough analysis of the resource model in online activism comes from two studies: Best and Krueger (2005) and Schlozman, Verba and Brady (2010). Both of these studies were based on surveys of on- and off-line political participation. Both studies find that online participation continues to be stratified by income and education—representational distortion is alive and well on the Internet. However, in a somewhat surprising reversal, Schlozman, Verba, and Brady show that age is negatively correlated with most online political acts. In a similar vein, Best and Krueger show that internet skills (designing webpages, using email) are predictive of online participation, but traditional civic skills (writing letters, chairing meetings) are not.

Broadly speaking, the evidence in these papers supports the classic "civic volunteerism model" of political participation (Verba, Schlozman and Brady, 1995). However, the papers focus primarily on low-intensity online political acts, such as sending an email, signing a petition, or donating money. Blogging is not addressed at all by Best and Krueger[5], and only in passing by Schlozman, Verba, and Brady. Fur-

[5] Their survey was fielded before political blogging came to prominence.

thermore, these studies seem designed to confirm existing theory, with small allowance for discovery of new features. All the main results concern the "average participant," with virtually no exploration of heterogeneity in participants' motives, ideology, or interests. Moreover, the studies barely engage with the qualities that make blogging unique, such as potential for advertising revenue, and competition for attention, traffic, and popularity. These studies include little discussion, and no data or analysis at all, about the content of online political participation.

The same criticisms could be applied to many other such studies as well. As a general rule, researchers investigating blogging through the lens of the resource model and its variants have contented themselves with replicating well-established findings from the study of political participation, sometimes with the addition of Internet-specific skills. This is worthwhile and important groundwork, but it misses the opportunity to extend and elaborate theory based on new information from online participation.

Stepping back, we can see that existing research into online political participation falls largely into one of two approaches. One approach focuses on thick descriptions of new media; the other, hypothesis tests of old theories. In this paper, I aim to fill a niche in the emerging literature on online political activism by combining the best of these two approaches. To date, these two branches of the literature have not fully engaged with each other.

2.3 Research Questions

If we accept the premise that political blogging is a form of activism, then there are at least three distinct directions we can pursue the relationship. First, we can explore the contours and dynamics of blogging in its own right. Second, we can compare and contrast blogging with other modes of activism. Third, we can investigate aspects of participation that were previously difficult to observe, but are easily measured in the blogosphere. To meet our goal of situating blogging within the Democratic repetoire of participation, we need to do a little bit of all three. This approach fills a niche in the literature on online political participation by integrating aspects of hypothesis testing and thick description. Hypothesis tests are valuable because they allow us to fit blogging into the framework of existing theory. Description is valuable because it can reveal directions for theory to grow and change.

Accordingly, this section introduces seven specific research questions to answer in the course of our analysis. These questions are framed in empirical terms, but each contributes to our theoretical understanding of blogging as an act of political participation. My hope is that these seven questions will collectively (1) add texture and depth to our exploration of the political blogosphere, (2) highlight areas where blogging may depart from traditional activism, and (3) provide the basis for discoveries that can inform future theories and models of activism.

With those considerations in mind, here are the seven questions.

1. How do the demographics of the political blogosphere differ from those of the electorate at large? Demographics provide an intuitive baseline for comparing bloggers to other populations of interest. Since demographics are also often strong proxies for other theoretically important variables, they can point us in the direction of noteworthy new patterns. In addition, socioeconomic status is traditionally the best predictor of representational distortion in politics, making these demographic variables the first place to look to understand whether blogging is likely to increase or decrease inequalities in participation and representation. Given past research into online activism, I hypothesize that bloggers will have more education and earn more income than U.S. adults on average, and are more likely to be white and male (Best and Krueger, 2005). Following Schlozman, Verba and Brady (2010), I also hypothesize that bloggers will be younger than other activists, but older than the population at large.

2. Are bloggers representative of the electorate in terms of ideology and partisanship? This question is important for understanding the impact of blogging on representation. Karpf (2009) and others have argued that the blogosphere leans left. More broadly, the popular stereotype of a political blogger is a relative extremist—a "wingnut" or "moon bat." If digital participation lends greater volume to progressive or extreme voices, we would certainly want to know it. With the first representative sample of political bloggers in hand, we are in a position to test these hypotheses for the first time.

3. To what extent do bloggers participate in offline politics? Understanding bloggers patterns of off-line behavior is an important step to understanding the role that blogging plays in modern activism. There are two competing viewpoints here: on one hand, critics (e.g. Keen, 2007) often paint bloggers as reclusive introverts, "activists" who only get involved from behind the safety of a computer screen. On the other hand, Reynolds (2007), Shirky (2009), and others contend that blogging is about communication and connection. According to this story, bloggers are typically mobilized within dense social networks, both on- and off-line. In many ways, this question parallels one of the ancillary questions asked by Verba, Schlozman and Brady (1995): are political participants specialists or generalists? Participation in offline modes of activism is one good way to judge between these two caricatures of bloggers.

4. Which aspects bloggers' of demographics, ideology, and political participation are correlated with popularity? Since "popularity" is unmeasured in most previous research, existing theories have little to say on this topic. It seems reasonable to suppose that popular bloggers are closer to political elites (Zaller, 1992), but there are at least two categories of elites to consider. If popular bloggers resemble professional journalists, we might expect them to be more centrist and politically neutral than other bloggers. Conversely, if popular bloggers resemble professional political operatives, then we would expect them to be more partisan and ideologically extreme. In either case, it seems reasonable to suppose that bloggers' education and income would increase with popularity, although I would expect the gains to be minor.

5. How much revenue and attention do bloggers report for their blogs? Traditionally, the two strongest incentives for joining a profession are money and prestige. Studies of political participation have largely focused on volunteer activism[6], and have therefore ignored these professional incentives. However, the structure of blogging enables even part-time bloggers to realize some of these rewards by building an audience and earning advertising revenue (Davis, 2009). If bloggers are influenced by these motivations, we must understand them in order to understand the shape and likely direction of participation in the blogosphere.

6. What reasons do bloggers themselves offer for political blogging? Asking bloggers directly about their reasons for blogging can shed a great deal of light on the goals they are pursuing and the constraints they face. From a qualitative perspective, this analysis can help us see if the answers bloggers give are similar to those we would expect from activists. Accordingly, we will replicate two batteries of questions about motivation from the landmark study by Verba, Schlozman and Brady (1995), as well as an additional battery of similar questions from the 2009 Technorati "state of the blogosphere" survey (White and Winn, 2009).

7. Does the resource model accurately describe patterns of participation in the political blogosphere? This line of inquiry calls for replication of classic

[6] A parallel tradition investigates quasi-professional activism, often focused on the opinions and demographics of local party members. See, for example, Katz and Eldersveld (1961), Jennings and Farah (1981), and Huckfeldt and Sprague (1992).

studies of participation in order to see whether time, money, skills, and political interest and efficacy are important predictors of blogging. Demonstrating this similarity would establish a deep theoretical connection between blogging and these other acts. Falsifying this claim would prove that engaging in political blogging is genuinely new and different, a critical step towards establishing online exceptionalism in theories of political behavioralism. As described previously, others have already provided suggestive evidence in favor of similarity, so I expect to find that blogging is driven by the same social forces that influence other acts in the participatory repertoire. My first aim for this question is to fill in the details and provide conclusive evidence, so that debate can move on. My second aim is to take stock of any important differences, because these will be key to understanding the new niche that blogging has come to occupy.

2.4 Methodology

In order to answer the research questions outlined in the previous section, we need to set up comparisons among several groups: bloggers, U.S. adults, and the subset of U.S. adults who are active in politics. In addition, we need to be able to compare bloggers across levels of popularity.

Our ability to set up these comparisons depends crucially on obtaining representative samples from both the U.S. population in general, and political bloggers in particular. Good methods for obtaining samples of the first type of been part of the practice of survey methods for many years. However, drawing a representative sample of the second type is only possible because of the methods developed in chapter I. Because we have a representative sample, rather than a convenience or prominence sample, we can draw conclusions about the blogosphere as a whole, not just popular A-list bloggers.

This section outlines my methodology for conducting surveys among representative samples each of these groups. First, I describe the parallel surveys that supply the data that makes up the backbone of our analysis. Second, I address important issues in measuring popularity among political bloggers. Finally, I describe methods for applying appropriate sample weights.

2.4.1 Parallel Surveys

Our analysis is based on surveys of two populations, which I describe here in turn.

One survey, called the Evaluations of Government and Society Survey (EGSS), was conducted around the 2010 midterm elections by the American National Election Study. Administered by the survey firm Knowledge Networks, the survey was conducted using an online self-administered questionnaire. Respondents were drawn from Knowledge Networks' respondent pool, a nationally representative panel of U.S. adults recruited by phone and maintained over time. The questionnaire included many common items from political surveys, as well as batteries of questions suggested and reviewed by the research community. In addition to the EGSS questionnaire, results from the Knowledge Networks general profile and Public Affairs profiles are also included in the data set. These profiling questionnaires were administered to respondents prior to the EGSS questionnaire, and include items about general demographics, as well as political attitudes and behavior.

I conducted the other survey shortly afterward. For notational convenience, I will refer to this survey by the name the survey team used internally: the Online Political Speech Project (OPSP) survey. Also a self-administered online survey, the OPSP survey was directed specifically towards political bloggers. Respondents were asked about their blogging habits and reasons for blogging, as well as standard batteries of political questions on voting, party identification, media consumption, attitudes towards groups and policies in society, and so on. Crucially, many of the items in the OPSP survey instrument were copied verbatim from the EGSS. This duplication enables us to compare results from one survey to the other. I also replicated items from Verba, Schlozman, and Brady's classic study of political activists (Verba, Schlozman and Brady, 1995), the 2009 Technorati "State of the Blogosphere" survey (White and Winn, 2009), and a few other items from the 2010 Cooperative Congressional Election Study.

The OPSP survey was administered to a representative sample of authors of active, English language blogs about U.S. politics. Although the technical details of constructing this sampling frame and contacting bloggers are quite involved, the gist is simple. First, I used web-crawling software guided by a highly accurate text classifier to conduct a census of English-language political web sites. From there, I drew a sample of sites, stratified by popularity, and filtered out those that were not blogs, not active, or not about U.S. politics. Finally, I gathered email addresses from the remaining sites and contacted bloggers with requests to participate in the study. Response rates among the contacted bloggers were reasonably high (around 25 percent), and post-hoc analysis suggests that non-response was largely random. We will address the potential for response bias in section 2.4.3. For further details on the survey administration, readers are referred to Appendix B for details. The full survey instrument is contained in Appendix C.

Because so much turns on the comparability of these two samples, it is important to stress the similarities between them. Both surveys were conducted using the same mode, a self-administered, online questionnaire. Both surveys shared many items in common: many question wordings and answer categories for the OPSP survey were copied directly from the EGSS. Finally, the surveys were conducted at around the same time. The EGSS was fielded in late October, shortly prior to the 2010 midterm elections. The Knowledge Networks general profile and Public Affairs profiles had been conducted previously, in most cases around February 2010. The OPSP survey was fielded in three waves, two pilot waves (total $n = 173$) in October and November

2010, and a second wave in May 2011 ($n = 603$). Since most of the variables that will be analyzed are essentially static among fully socialized adults, the passing of a few months is unlikely to make much difference.

2.4.2 Measuring popularity

Since a substantial portion of our analysis will be dedicated to understanding differences among bloggers by popularity, it is important that we have accurate measurements for popularity. By *popularity*, I mean "the total amount of attention that the blog receives." Within our data set, we have three possible measures of popularity to choose from: page views, unique visitors, and inbound links. This section describes each of these variables briefly, and then explains why I selected page views as the primary measure of popularity.

Page views are the most direct measure of site traffic. Each time a user clicks on a link or otherwise directs her browser to a URL within a given site, the site's servers record a single page view. Clicks within the site also count as page views: it doesn't matter whether the traffic originates externally or internally. Generally speaking, page views are private information known only to a site's administrators. To obtain this information, our survey asked each blogger, "About how many page views did your blog receive last month?"

Unique visitors is a somewhat messier metric. In theory, it measures the number of people who visited the site within a given timeframe. Unlike page views, a user who visits multiple pages within the same site should only be counted once. Moreover, a visitor who leaves the site and returns later should still be counted only once. However, a sites' ability to track users across sessions depends on several outside factors such as browser settings and third-party cookies, and is rarely 100 percent reliable. We measured unique visitors through another survey question: "About how many unique visitors did your blog receive last month?"[7]

Our measure of *inbound links* comes from our census of political web sites. It is the number of hyperlinks direction to the site in question, from other sites visited in the course of the snowball census. Past research has demonstrated that inbound links are correlated with site traffic, especially among top web sites: Hindman, Tsioutsiouliklis and Johnson (2003) found a .71 correlation between inbound links and site traffic as measured by the web-tracking company, Alexa. As a measure of popularity, inbound links have the unique virtue of being based on objective data, not just bloggers' self-reported statistics. However, links are only a proxy for the total attention that a blog receives.

Which of these measures should we use in our analysis? Ideally, we would prefer to use page views, because it is has the best theoretical motivation. But since the variable is self-reported, we should take care to validate it carefully first.

As a first step for validation, I checked for outliers, finding two. The first site claimed exactly 1 billion monthly page views, and the second claimed an even larger number: 1.828494e+48. A quick web search showed that neither site is a popular, well-known blog. Given that the next most popular site claimed less than 2 million monthly page views, I felt safe excluding these two from analysis.

As a second step for validation, I replicated Hindman, et. al's analysis by comparing correlations between all three potential measures of popularity. Our primary goal in this analysis was to check our (self-reported) measure of blog popularity against our (objective) measure of inbound links. A secondary goal was to understand the relationship between self-reported page views and unique visitors. For all blogs in the sample, page views and unique visitors are tightly correlated (r = .90), and each is weakly correlated with inbound links (r = .24 and .26, respectively.) For sites with at

[7] In retrospect, this question wording was probably a mistake. Most web analytics report unique users by day, not by month. In comments on the survey, a couple bloggers mentioned being confused by this difference. Judging by the numbers, many others probably reported unique daily visits, not monthly visits. These issues further cloud an already complicated measure.

least a thousand inbound links (n=68), the correlations are considerably stronger .997, .620, and .593, respectively. For the 17 sites with over 2,000 inlinks, the correlations are all above .99. These findings validate our self-reported measure for page views against a credible, objective metric. In addition, the patterns of correlation match those of Hindman, et. al. quite closely

Taken together, these facts are persuasive evidence that user-reported page views are a reasonable measure of real web traffic. Accordingly, I adopted self-reported page views as the primary measure of blog popularity for all analysis in this chapter. As a robustness check, I also replicated all analysis using unique visitors.

2.4.3 Sample weighting

From the perspective of sampling theory, our research questions call for three different types of analysis, each of which requires a different approach to sample weighting. Accordingly, I used two types of sample weights: design and response weights. This section begins by describing these two types, then proceeds to explain which weights were applied in which analysis.

The first type of sample weight is *design weights* reflecting our sampling strata. In order to guarantee a sufficiently large sample of "A-list" bloggers, I deliberately oversampled popular bloggers, based on inbound links[8]. This stratification was essential for drawing comparisons among political bloggers by levels of popularity. However, it also makes it is necessary to down-weight cases from the upper strata when computing statistics for the average blogger. Because of the long tail in blog popularity, our sampling proportions are quite unequal, leading to unusually severe downweighting. Bloggers in the most heavily oversampled stratum were roughly 6.5 times more likely to be sampled than bloggers in the next stratum, and 108 times

[8] For sample stratification, bloggers' responses were unavailable, of course. Both my and Hindman et. al.'s analysis show that inbound links are an effective way to identify very popular web sites, so this application is reasonable for our purposes.

more likely than those in the final stratum, leading to sampling weights of .009, .060, and 1.000, respectively.

The second type of sample weight is *response weights* intended to correct for the possibility of selective nonresponse (Groves et al., 2002)[9]. I accomplished this with a propensity score approach: using data available for both respondents and nonrespondents, I modeled the probability of individual-level responses, and re-weighted the sample to give more emphasis to bloggers of underrepresented types. The net result is population-level estimates that are more reflective (in expectation) of the underlying population (Rosenbaum and Rubin, 1983).

In theory, many surveys would benefit from response weights estimated via propensity scores (Dehejia and Wahba, 2002). The main reason that propensity scores are not used more often is that researchers typically lack data on the non-responding population. In this case, studying the blogosphere gives us a bit of an advantage: we can use the contents of blogs to generate propensity scores and sample weights, using the following procedure. First, for each blog in the intended sample (both respondents and non-respondents), crawl the homepage of the blog. Second, tokenize the page and extract presence vectors for all common features[10]. Third, calculate propensity scores by regressing blog features onto survey nonresponse in a binary logistic model[11]. Finally, take the inverse of the propensity score as the response weight on each blog.

This approach to sample weighting is unconventional, but follows the familiar logic of propensity scores[12]. The key assumption is that any biasing variables leaves

[9] In addressing the issue of onresponse, we have one important advantage: our target population universally has access to the Internet. This is one of the most important threats to validity in many Web-based surveys (Couper, 2000). Still, the causes of nonresponse in Internet surveys are not fully understood. See Appendix B for a description of the procedures I used to maximize response rates.

[10] I accepted all alphanumeric strings as valid tokens, and retained the 6,000 most common tokens for use in the model.

[11] After some experimentation, I used L1 regression with a .01 regularization constant. This model had a modest pseudo-R-squared of .12. To avoid overfitting, propensity scores were calculated in the style of a 25-fold cross-validation: in each iteration, 24/25ths of the sample was used to train the model, and propensity scores were generated for the remaining 25th. Propensity scores for respondents and nonrespondents showed sizable overlap between groups.

[12] Actually, it is not so different from propensity score matching as used to maintain web panels, as

an imprint on the text of the blog. Unlike the problem of unbalanced stratification described previously, weights assigned via propensity scores are not so dramatically unbalanced. Therefore, they do not have as much of a negative impact on effective sample sizes.

Having discussed these types of sampling weights, we now turn to types of analysis. This chapter contains three major types: population averages, comparisons of bloggers by strata, and comparisons of multivariate partial effects.

First, several of our research questions call for comparisons of *population averages*: typically, the average blogger versus the average U.S. adult. In this case, we want our mean statistics and surrounding confidence intervals to represent the population as accurately as possible, without bias. Therefore I apply both sampling and response weights[13].

Second, a number of our research questions concern *comparisons of bloggers across strata*. For questions of this type, I used response weights but not design weights. This approach is appropriate, because we wish to correct for nonresponse bias, but preserve sample stratification in order to draw comparisons across popularity levels.

Finally, the final piece of analysis requires *comparisons of multivariate partial effects*. This analysis tests the resource model in the context of the blogosphere. For this analysis, I use unweighted data, for three reasons. First, the analysis hinges on comparisons of multivariate partial effects. Unlike comparisons of group averages, it is not clear whether applying sampling weights will yield more accurate estimates. Second, given the presence of many correlated variables in the regression models, the penalty for sample weighting on effective sample size becomes more problematic. In other analysis, we can afford to use conservative weighting schemes, even at the cost of inflated standard errors, because sample sizes of a few hundred are still sufficient to

in Lee (2006) and Rivers (2007). The main difference is using text as covariates, rather than previous survey resposnes, to correct for non-response and attrition.

[13] I created combined weights by multiplying design weights and response weights. This approach assumes that nonresponse is unconditional (or approximately unconditional) on strata.

yield conclusive results. The same is not true in this multuvariate analysis. Finally, and perhaps most importantly, our primary hypotheses for this analysis focus on differences between blogging and other forms of activism. Adopting overly conservative methods, and erring on the side of false negatives can potentially confuse our results even more than small differences due to lack of sample weights.

A final thought on sample weighting: it should be mentioned that none of these weights has much effect on the substantive conclusions of the chapter. There are small differences in nuance, of course, but the major results seem to hold up whether we use design weights, response weights, both, or neither. As a rule, sample weights yield larger standard errors, and make our estimates more conservative, but do not affect the overall patterns in the data.

2.5 Imputing civic skills

Results for our final research question require measures of civic skills for both bloggers and U.S. adults. Following classic participation studies (e.g. Verba, Schlozman and Brady 1995) and more recent replications (e.g. Burns, Kinder and Ortiz 2002), the OPSP survey included several complementary measures of civic skills acquired through organizational participation: "In the last six months, have you given a presentation or speech in the following settings?", and "In the last six months, have you planned or chaired a meeting in the following settings?" Settings included "At work," "At your church or place of worship," "In some other organization."

Unfortunately, the EGSS did not include comparable measures. To compensate for that shortfall, I imputed values for civic skills with a two-stage auxilliary instrumental variables (2SAIV) model (Franklin, 1989). Using data from the OPSP survey, I estimated parameters for a simple regression model of civic skills based on measures of news consumption, trust in government, church attendance, and political discussion in various settings. Since each of these variables is included in both surveys, I was able to generate imputed values for civic skills among both U.S. adults and bloggers. To maintain parity, we then use the imputed variables as our measures of civic skills within both data sets.

As discussed by Franklin (1989), the key assumption for this procedure is that the relationship between civic skills and its predictors (e.g. news consumption, trust in government) is the same between bloggers and U.S. adults. This is an admittedly strong assumption, but the limits of the EGSS data make 2SAIV the only feasible approach for this estimation. Moreover, as seen in column one of Table 2.2, results from the procedure are quite reasonable. Consequently, in the spirit of making the best use of available data, I report the 2SAIV results even though the procedure depends on strong statistical assumptions.

2.6 Results

We now turn to results. This section presents answers to our seven specific research questions. For the most part, this section describes the outcome of analysis without interpretation. Discussion is saved for Section 2.7.

2.6.1 Demographics

Demographic comparisons of bloggers and the U.S. population reveal some striking differences. Over 80 percent of the political blogosphere is male, and 62.7 percent is white[14]. Oddly, 1 in 4 bloggers do not identify as white, black, or Hispanic. The open-ended "other" category included several bloggers who declined to state an ethnicity, some multiracial bloggers, and at least two "humans."

Bloggers tend to have much more formal education and substantially higher income

[14] A note on standard errors and confidence intervals: all standard errors in this section are based on 5,000 bootstrapped iterations with appropriate weights; unless otherwise noted, dotplots and tables report 95% confidence intervals, based on the same bootstrapping approach. As with weighting, bootstrapped intervals had little impact on substantive results.

than the rest of the population. The median blogger has a Bachelors' degree and has a household income over \$60,000 a year. As many as 48.1 percent of bloggers have advanced degrees, compared to 10.9 percent in the general population.

Somewhat surprisingly, the mean age (45.6) among political bloggers is almost exactly the same as the population average (49.1). Comparing the distribution of ages, we find that political bloggers are concentrated between the ages of 30 and 60. There are few very young or very old bloggers.

Table 2.1 reports comparative proportions for all of these demographic categories. Chi-squared tests reveal that all of these differences are significant with $p < .0001$.

2.6.2 Partisanship and ideology

In terms of partisanship, bloggers are somewhat less likely to identify as Republicans and Democrats. Independents and "others" comprise 32 percent of the blogger sample, compared to 18.2 percent in the EGSS. This pattern is more pronounced on the right side of the aisle: only 14.6 percent of bloggers self-identify as Republican, compared to 26.6 percent in the general population. Several bloggers who self-identified as "other" volunteered that they consider themselves libertarians. On the whole, bloggers seem reluctant to identify with either major party.

This pattern is somewhat at odds with bloggers' self-identification on the liberal-conservative scale. When asked about ideology, bloggers are much more likely to identify as relative extremists. In particular, the "extremely liberal" category includes almost a quarter of political bloggers, compared to only 6 percent of the population at large. We see the same trend in reverse at the middle of the distribution: nearly a third of respondents on the EGSS said they were "neither liberal nor conservative," compared to only 16 percent of the blogger sample. Taking these results at face value, it most political bloggers express stronger affinity for ideology than party.

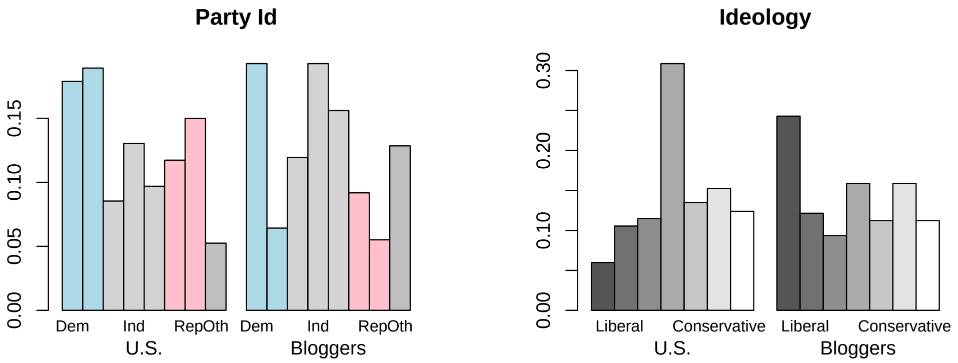
In terms of balance, bloggers appear to be reasonably representative of the pop-

Table 2.1: Demographics among U.S. adults and bloggers

Variable	Categories	U.S.	Bloggers
Gender	Male	48.2	84.6
	Female	51.7	15.3
Ethnicity	White	72.5	62.7
	Black	11.7	3.7
	Hispanic	10.0	3.9
	Other	5.6	29.6
Age	18-19	2.5	0.0
	20's	18.8	13.9
	30's	16.1	23.1
	40's	17.9	20.2
	50's	18.8	26.8
	60's	13.3	10.2
	70+	12.3	10.2
Education	Less than high school	13.2	0.0
	High school/GED	29.2	3.6
	Some college	18.6	12.3
	Associates degree	7.5	8.3
	Bachelors degree	20.2	27.3
	Masters degree	8.3	17.7
	Professional or PhD	2.6	30.4
Income	Less than $10k	9.9	2.0
	$10k - $19k	13.5	7.4
	$20k - $29k	11.5	6.1
	$30k - $39k	15	6.9
	$40k - $49k	8.9	8.7
	$50k - $59k	10.7	7.8
	$60k - $99k	21.6	21.0
	$100k - $149k	5.3	24.0
	More than $150k	3.2	15.6

Statistics are weighted means from the EGSS and OPSP surveys, collected around the 2010 midterm elections, with sample sizes of 1,240 and 776, respectively. See Appendix B for details.

Figure 2.1: Party ID and ideology among U.S. adults and bloggers



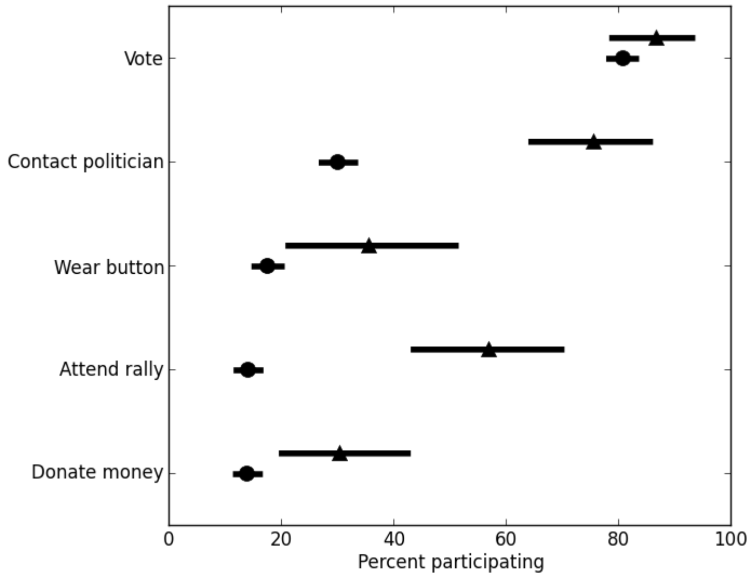
These graphs show partisan and ideological distributions for the weighted EGSS and OPSP surveys. Data were collected around the 2010 midterm elections, with sample sizes of 1,240 and 776, for U.S. adults and bloggers, respectively. See Appendix B for details.

ulation at large. If we classify leaners with their parties, we find that 37.6 percent of bloggers favor the Democratic party and 30.2 percent favor the Republican party, compared to 45.3 and 36.4 percent, respectively. By these numbers, 55.4 percent of partisans in both populations support Democratic party. In ideology, we find evidence of moderate overrepresentation of liberals in the blogosphere: 54.4 percent, instead of 40.5 percent. Figure 2.1 illustrates distributions for party ID and ideology for bloggers and the general population.

2.6.3 Political engagement

Earlier, we raised the question whether bloggers are "loners," or involved in other modes of political participation as well. The answer is unambiguous: the typical blogger is a frequent participant in other political venues. Three out of four political bloggers (75.5%) report phoning, e-mailing, writing to, or visiting a government official in the last year. Nearly sixty percent (56.9%) attended a political speech, rally, or demonstration. More than a third of bloggers (35.4%) advertised for campaigns, and 30.4 percent gave money to political causes. As shown in Figure 2.2, comparable statistics for the US population were all 25 percent or lower. Bloggers are politically

Figure 2.2: Political participation among U.S. adults and bloggers



This figure is a comparative dot plot showing average political participation among bloggers and the U.S. population as a whole. Each of four political acts is shown as a pair of rows. On the upper row, the percent of political bloggers who engaged in that form of participation is marked by a triangular dot. On the lower row, the percent of U.S. adults is marked by a round dot. In each row, 95 percent confidence intervals are shown by horizontal bars.

active off-line as well as online.

Bloggers' other attitudes are consistent with this pattern of political engagement. Bloggers are very interested in politics, with 92.0 percent saying they are "very" or "extremely" interested in government and politics. The population baseline is 40.2 percent. Bloggers also have higher than average political self-efficacy, with 35.6 percent of bloggers answering "a great deal" or "a lot" to the question "How much can people like you affect what the government does?" Among U.S. adults, the statistic is 21.4 percent.

2.6.4 Differences in demographics, ideology, and participation by blogger popularity

This question is easily answered: there are virtually no important differences in demographics, ideology, or participation by blogger popularity. In a bivariate regression, gender is slightly correlated with logged page views ($\beta = -.030; t = -2.41$). Loosely speaking, we can interpret the coeficient as "a 100 percent increase in blog traffic is associated with a 3 percent increase in likelihood that the author is male." Aside from this small effect, none of our other demographic variables are correlated with our measure of popularity.

The same null result holds true for partisanship, ideology, ideological extremism (i.e. distance from the midpoint on a seven-point ideological scale), interest in politics, political efficacy, and participation in each of the modes described in the previous section. This conclusion was frankly surprising, but it holds up under multiple specifications and related variables for popularity. Evidently, popularity in the blogosphere is essentially uncorrelated to demographics, ideology, and offline participation.

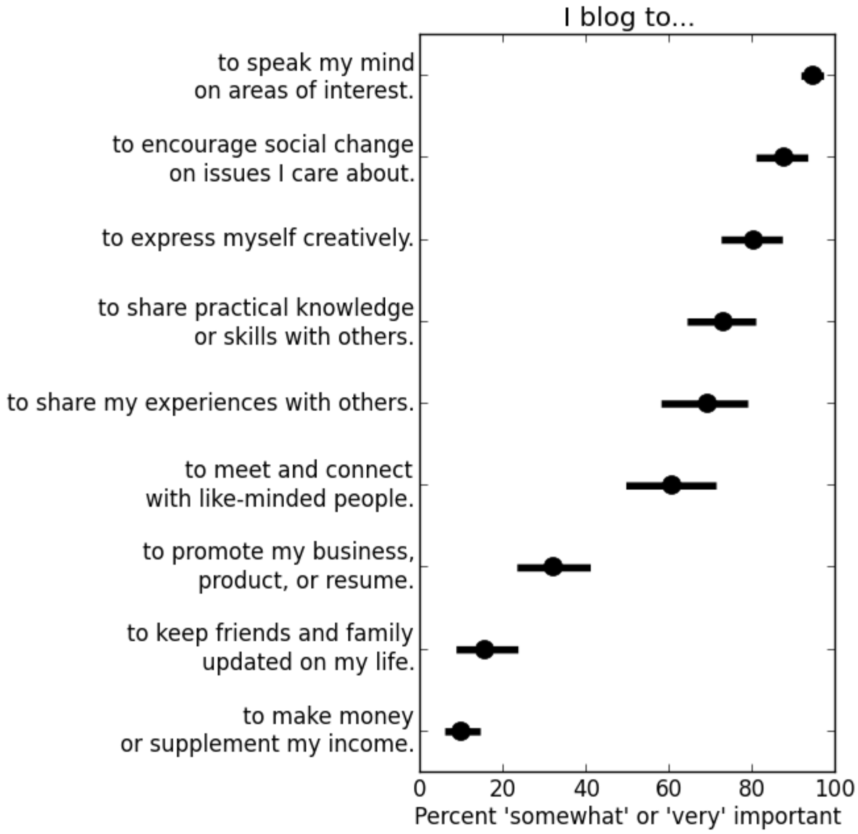
2.6.5 Self-reported reasons for blogging

One way to learn why people blog about politics is to ask them directly. Direct, self-reported reasons can help us understand the culture of political blogging: the stories bloggers tell each other—and themselves—about why they blog. They can also help us understand important differences among political bloggers.

The blogger survey included two batteries of questions about motivations for blogging. The first battery was based on the 2009 Technorati "State of the Blogosphere" survey. Respondents were asked how important each of several factors is in their decision to maintain a blog. Of these, the most popular answers were "I blog to speak my mind on areas of interest," with 95.3% "somewhat" or "very important" (Reported percentages in the next four paragraphs are all for the same answer categories) and "I blog to encourage social change on issues I care about" (87.4%). Figure 2.3 reports results from all the items in this battery.

The second battery of questions was replicated from the 1990 study of political

Figure 2.3: Self-reported reasons for blogging: Technorati questions



This dot plot shows the percent of bloggers who agree with each statement, estimated using sample weights across the OPSP survey (n=776). Bootstrapped 95 percent confidence intervals are shown by horizontal bars.

activists conducted by Verba, Schlozman and Brady (1995). Bloggers were asked to respond to 15 statements about reasons for blogging[15]. Of these, the most popular was "When I blog about politics, I do it for the chance to make the community or nation a better place to live" (85.1%), followed by "for the chance to work with people who share my ideals" (73.8%), "because it's my duty as a citizen" (68.6%), "for the chance to influence government policy" (68.5%), "to learn about politics and government" (65.3%), and "because I find it exciting" (57.3%). Figure 2.4 reports results from all the items in this battery.

Through these questions, political bloggers express a wide array of motivations. Many bloggers claim to be intent on policy change, as evidenced by the "encourage social change" and "influence government policy" questions. Large numbers of bloggers also cite social (e.g. "to meet and connect with like-minded people," "to work with people who share my ideals," "to be with people I enjoy,"), emotional ("to express myself creatively,"), and civic ("to learn about politics and government," "because it's my duty as a citizen") reasons as their key motivations for blogging. Few (but not zero) bloggers listed particularized benefits ("to promote my business, product, or resume," "to further my job or career," "to get help from an official") as reasons for blogging about politics.

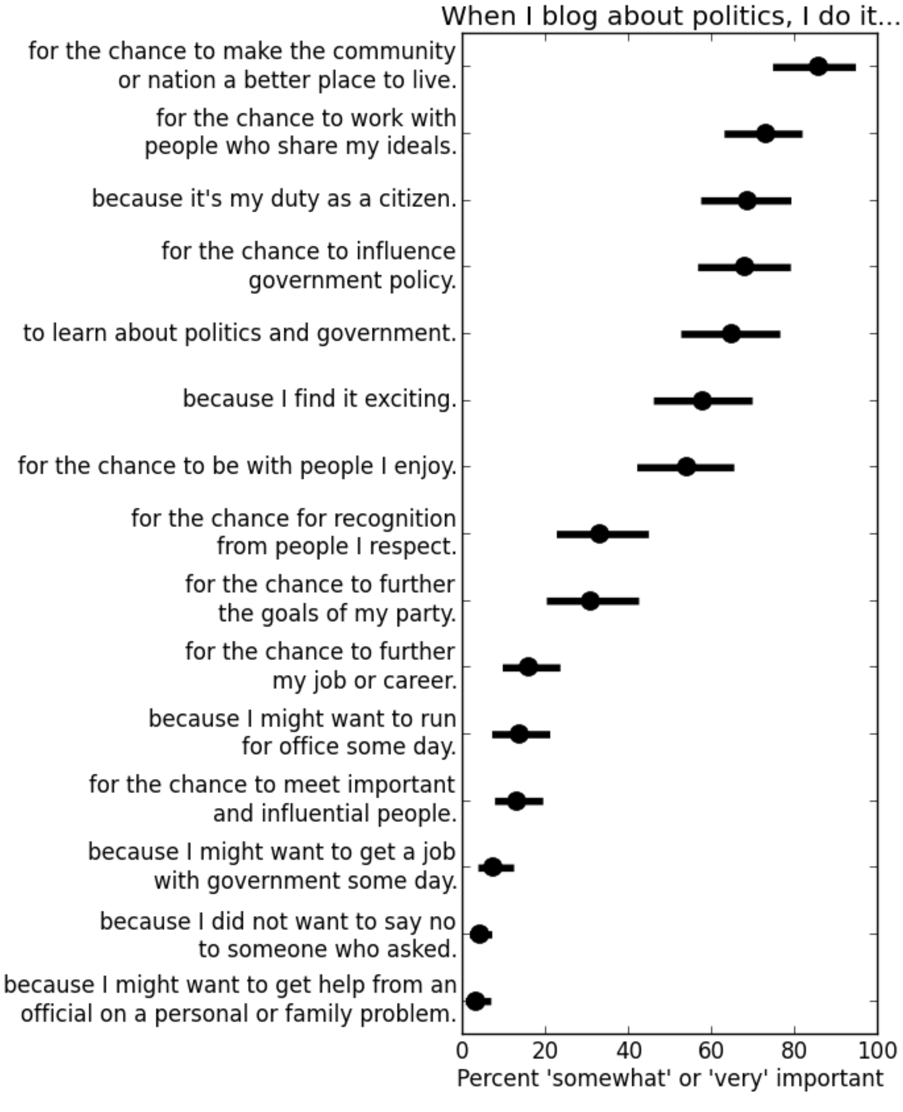
Because they are self-reported, these answers should be taken with a grain of salt. Most likely, they reflect bloggers' internal justifications for participation in the political blogosphere. They don't necessarily reflect all the underlying causes and motivations for blogging. Still, these answers provide corroborating evidence for the "blogging as activism" thesis: they closely resemble the kinds of explanations we would expect from political activists.

Interestingly, none of the motivations seems to be correlated with popularity. Three of the survey items[16] show small correlations (two negative, one positive) with

[15] Questions were reworded slightly to accommodate the self-administered survey format.

[16] The three items are "to share my experiences with others," "to make money or supplement my

Figure 2.4: Self-reported reasons for blogging, questions from Verba, Schlozman, and Brady



This dot plot shows the percent of bloggers who agree with each statement, estimated using sample weights across the OPSP survey (n=776). Bootstrapped 95 percent confidence intervals are shown by horizontal bars.

page views, but the relationships barely cross the p < .1 threshold for statistical significance. When I attempted to replicate the analysis using unique visitors, none of the correlations was statistically significant. I conclude that bloggers' self-expressed motivations are largely unrelated to their popularity.

2.6.6 Revenue and recognition

Having investigated structure and goals, let us turn now to two of the most tangible rewards of blogging: revenue and recognition. As noted earlier, most bloggers are unpaid amateurs. However, a significant fraction do earn money from their blogs. Figure 2.5 shows the portion of bloggers who earn money from various sources. By far the most prevalent source of income is advertising, with 13.9 percent of bloggers making money through ads. Other sources of income from blogging include (in decreasing order of prevalance), outside employment (8.4%), tip jars (3.2%), selling items through the blog (2.3%), and premium content (less than 1 percent; only three bloggers in the sample.) There is some overlap between these categories: collectively 12.7 percent of bloggers earn money through means other than advertising, and 22.6 percent total of bloggers earn money by any of these means.

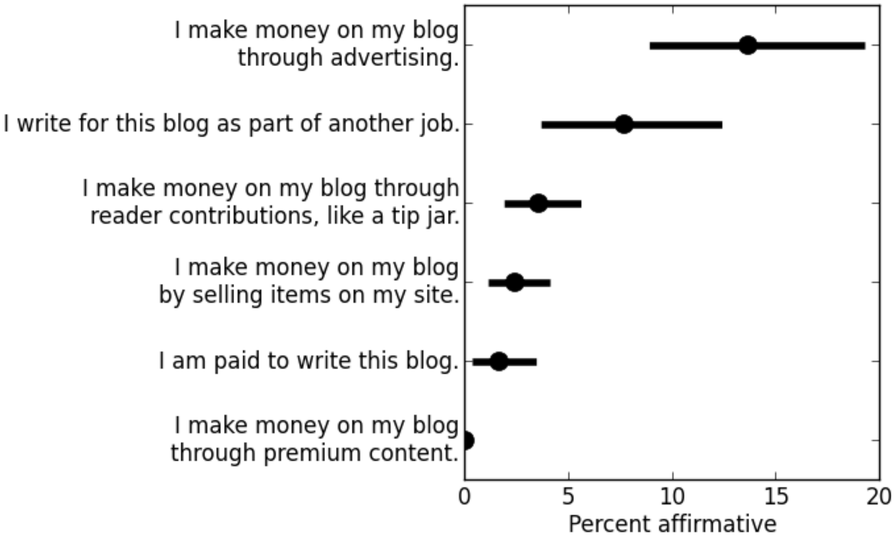
Among bloggers earning revenue, the distribution of revenue is highly uneven, as shown in Figure 2.6. Of political bloggers who made any money at all through their blog, more than a third (34.7%) made less than $50 last year, and roughly two-thirds (63.9%) made less than $1,000. However, a significant fraction of bloggers (18.2%) made more than $2,000 by blogging last year. Given that some of our respondents were full-time professionals, it seems reasonable to believe that their earnings amounted to at least tens of thousands of dollars[17].

Political bloggers' self-reported *advertising* revenues closely reflect *total* revenue,

income," and "for the chance to influence government policy."

[17] Unfortunately, I didn't anticipate the number of full-time bloggers who would respond to the survey, and so the answer scale does not go high enough to capture their responses in any detail.

Figure 2.5: Revenue sources for political bloggers



This dot plot shows the percent of bloggers who reported revenue agree with each statement, estimated using sample weights across the OPSP survey (n=776). Bootstrapped 95 percent confidence intervals are shown by horizontal bars.

suggesting that ads make up the lion's share of income for bloggers. All told, 88.8% of bloggers with any revenue made at least some of that money through ads.

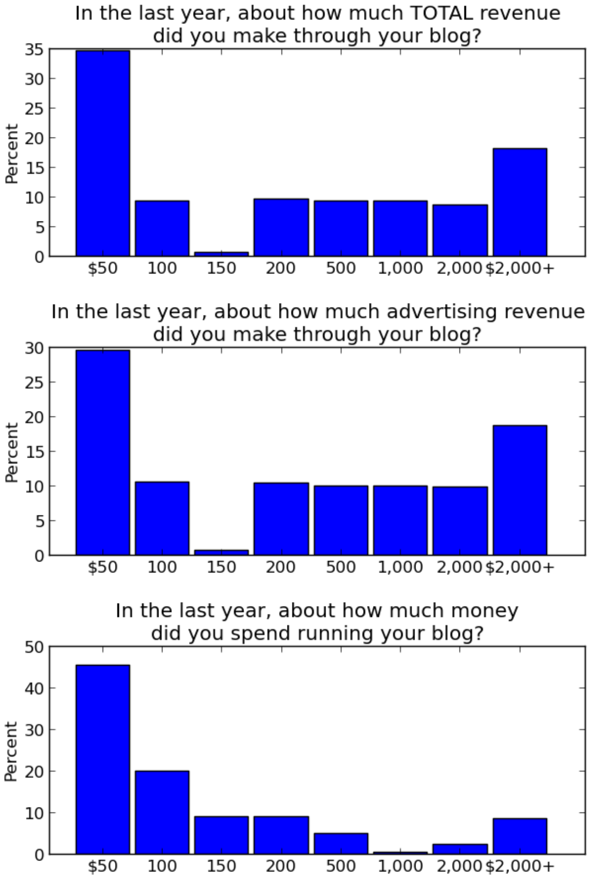
Political bloggers also incur expenses running their blogs, but these expenses tend to be small. Only 51.8 percent of bloggers reported spending any money running their blog last year. Of these, 29.5 percent spent less than \$50 and 61.3 percent spent \$500 or less. A small handful of bloggers (18.7%) reported expenses exceeding \$2,000. Intriguingly, the number of bloggers losing money on their blogs (29.9%) far exceeds the number that we can unambiguously determine are making money (5.7%).

Not surprisingly, popular bloggers are more likely to report earning revenue on their blogs. Of the six possibilities for generating revenue, advertising is the only category that shows a significant difference by levels of popularity ($\beta = 0.0244; t = 2.45$)[18]. Popular bloggers are also more likely to report higher figures for total revenue, ad revenue, and expenditures[19].

[18] Once again, I used logged page views to measure popularity.

[19] Unfortunately, the lack of an interval-level answer scale leaves us unable to estimate dollar amounts.

Figure 2.6: Ad revenue, total revenue, and expenditures among bloggers



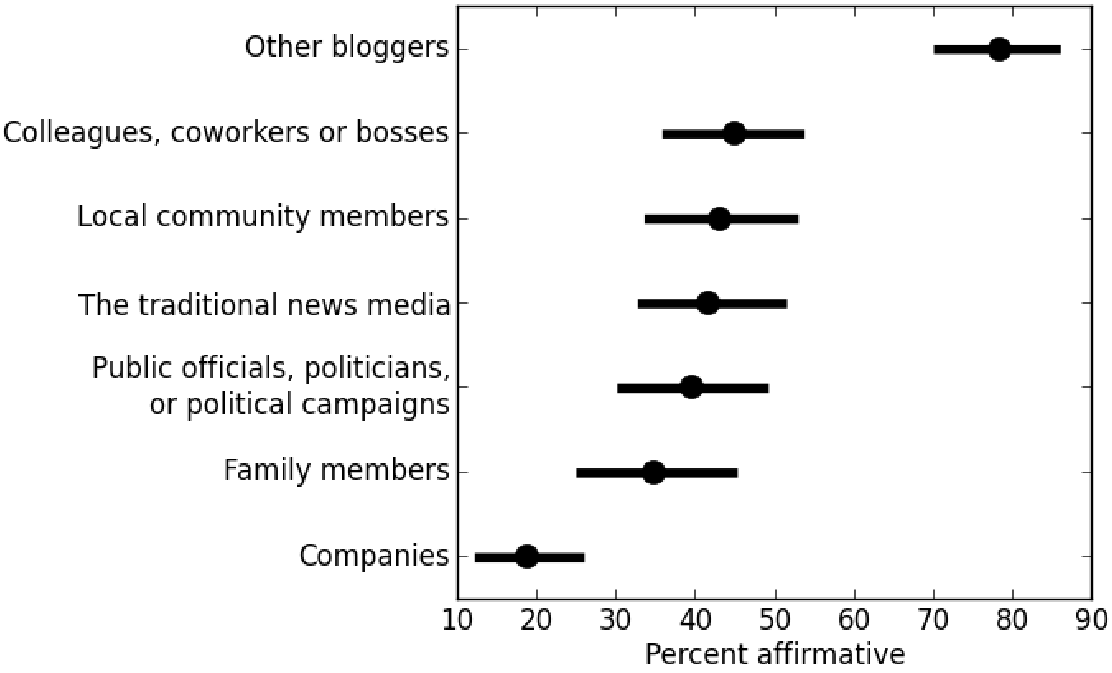
In the last year, about how much TOTAL revenue did you make through your blog?

In the last year, about how much advertising revenue did you make through your blog?

In the last year, about how much money did you spend running your blog?

Values are based on the weighted OPSP sample (n=776).

Figure 2.7: Sources of attention for political blogs



Values are based on weighted averages from the OPSP survey (n=776). Bootstrapped 95 percent confidence intervals are shown by horizontal bars.

We saw earlier that many bloggers post their views with the goal of influencing others. In contrast to the relatively small number of political bloggers making money, large majorities of political bloggers report that their blog has "received attention from or been mentioned by" a variety of people, as shown in Figure 2.7. Far and away the most important source of attention for political bloggers is other bloggers. Fully three out of four bloggers (78.2%) report interactions with other bloggers last year. Interestingly, mentions by other blogs seem to be a gating factor for mentions by other sources: only 3.6 percent of bloggers received attention from other sources without receiving attention from other bloggers.

From the perspective of policymaking, two other categories are particularly noteworthy. First, 41.7 percent of bloggers report attention from "traditional news media." Second, 40.1 percent report attention from "public officials, politicians or campaigns." These numbers are surprisingly high. In addition, these figures are correlated with popularity, with a 2 to 4 percentage-point increase in likelihood of attention each type of source for each doubling in monthly page views. Even accepting that bloggers may be optimistically overreporting their own influence, this level of contact with influencers and policymakers is impressive.

2.6.7 Replicating the resource model in the blogosphere

As a second way to look at the data, we can compare political bloggers, political activists, and U.S. adults in general. This approach enables us to move beyond most of the limitations of self-reported data. Here, I classify as an activist anyone who participates in at least two civic acts from the following list: voting, donating money to a political cause or campaign, advertising (e.g. wearing a button or posting a yard sign), attending a political rally or event, or contacting a government official. I will use the term "others" to refer to those who are not bloggers or activists.

Simple descriptive statistics and pairwise t-tests suffice for this analysis. Both bloggers and activists are more interested in politics than others (35.4% somewhat or very interested), with bloggers (89.1%) even more interested than activists (78.3%), and all differences statistically significant at the .01 level. Responding to the question, "How much can people like you affect what the government does?" activists reported higher political efficacy (76.7% answering "a moderate," "a lot," or "a great deal") than bloggers (63.6%, t=3.39), who were higher in turn than others (47.0%, t=6.99). Bloggers are more likely to be wealthy, with 61.9 percent earning over $60,000 per year, compared to activists (52.3%, t=2.35) and others (37.6%, t=3.82). Bloggers are also much more likely to hold an advanced degree (81.5%) than activists (46.5%, t=10.27) and others (29.1%, t=24.52). Bloggers are more likely to be male (77.3%) than activists (51.3%, t=7.25) or others (47.1%, t=13.08).

The relationship between blogging, activism, and age is somewhat more complicated. On average, activists are about 8 years older than others (t=6.57), and bloggers are about 2 years younger than activists (t=2.80). However, political bloggers are also underrepresented in the under 30 category. Thus, the overall pattern for age among political bloggers can be described as "younger, but not too young."

Overall, bloggers fit the demographic profile of activists quite nicely. With the exception of age, in every category where activists differ from the population at large, bloggers also differ in the same direction. These head-to-head comparisons help us understand the general similarities and differences among bloggers, activists, and the population at large. However, they cannot tell us which factors have the largest effects, conditional on the others.

To unpack partial effects, I set out to replicate Verba, Schlozman, and Brady's civic volunteerism model of participation (Verba, Schlozman and Brady, 1995) in the blogosphere. In order to mirror the original study as closely as possible, I included variables for political interest and efficacy, civic skills acquired through organizational affiliation, income, education, and age. Since not every variable of interest was available in the EGSS study, I followed the lead of Burns, Kinder, and Ortiz' careful replication study, and included several demographic variables: work status, an indicator variable for age greater than 65, marital status, race, and gender (Burns, Kinder and Ortiz, 2002). Because we understand the causal structure behind political participation reasonably well, these demographics can help serve as proxies for quantities that were not adequately measured in the EGSS.

Table 2.2 reports results from this model of participation applied to seven different measures of political participation[20]. In the first column, the dependent variable is the sum of participatory acts for respondents in the EGSS sample. These results serve as a sanity check: as expected, political interest and efficacy, organizational participation, income, education, and age are all positively correlated with one's overall level of political participation. These results closely replicate those of Burns, Kinder, and Ortiz. Despite the missing variables, they also capture the important theoretical results from the original resource model.

Column 2 in Table 2.2 reports results from a model that has the same covariates, but takes political blogging as the dependent variable. This model combines the EGSS and OPSP samples in a case-control design. It provides the primary test of several of our important hypotheses. Implicitly, we are comparing differences between the OPSP sample of political bloggers and the EGSS sample of U.S. adults, in order to see which characteristics are most strongly associated with political blogging[21].

Most of the broad strokes of this analysis are consistent with what we already know about political participation, but there are some surprises. Political interest is a strong,

[20] Note that the first model differs slightly due to the structure of the dependent variables. In order to fit a counted DV, the first model is a Poisson GLM. The other models are all logistic regressions, with binary dependent variables. Therefore, we can compare the direction and significance, but not the exact magnitudes, of coefficients from the first model against the others.

[21] This analysis assumes that all respondents in the OPSP sample are political bloggers (a safe assumption that was thoroughly verified in the sample construction process), and that none of the respondents in the EGSS sample are political bloggers. This second assumption might be violated, but only slightly. Based on our earlier estimate that one in 500 adults is a political blogger, we would expect no more than a small handful of bloggers in the EGSS sample. Moreover, any measurement error due to miscounting these cases is likely to attenuate regression estimates toward zero.

Table 2.2: Resource models of participation in blogging and other types of activism

	All acts	Blog	Vote	Donate	Advertise	Contact	Attend
(Intercept)	-1.46^{***}	-7.39^{***}	-3.01^{***}	-7.61^{***}	-5.29^{***}	-5.32^{***}	-5.45^{***}
	(0.13)	(0.50)	(0.41)	(0.71)	(0.47)	(0.44)	(0.53)
Pol. interest	1.03^{***}	5.19^{***}	2.01^{***}	1.96^{***}	1.88^{***}	2.93^{***}	3.87^{***}
	(0.11)	(0.46)	(0.39)	(0.53)	(0.42)	(0.37)	(0.53)
Pol. efficacy	0.35^{***}	-0.03	1.34^{**}	1.12^{**}	1.36^{***}	0.53	0.44
	(0.10)	(0.29)	(0.41)	(0.40)	(0.35)	(0.30)	(0.37)
Civic skills	0.23^{***}	0.18	0.58^{***}	0.64^{*}	0.80^{***}	0.66^{***}	0.73^{**}
	(0.05)	(0.15)	(0.35)	(0.24)	(0.21)	(0.20)	(0.22)
Income	0.42^{***}	0.46	1.49^{***}	1.75^{***}	0.57	0.55	1.56^{***}
	(0.11)	(0.31)	(0.41)	(0.52)	(0.40)	(0.36)	(0.49)
Education	0.28^{**}	3.65^{***}	1.04^{***}	0.28^{*}	-0.45	0.32^{**}	-0.77
	(0.09)	(0.40)	(0.40)	(0.41)	(0.36)	(0.33)	(0.39)
Working	-0.10	0.03	-0.11	-0.08	-0.26	-0.57^{**}	$-0.61^{\dagger}$
	(0.06)	(0.20)	(0.20)	(0.29)	(0.24)	(0.21)	(0.28)
Age	0.54^{**}	-0.94^{*}	2.72^{***}	4.20^{***}	1.43	$1.61^{\dagger}$	$-0.52^{\dagger}$
	(0.19)	(0.58)	(0.65)	(0.96)	(0.71)	(0.63)	(0.80)
Over 65	0.11	-1.09^{***}	0.58	0.12	0.38	-0.11	$0.51^{\dagger}$
	(0.08)	(0.32)	(0.20)	(0.36)	(0.31)	(0.27)	(0.35)
Married	0.05	-0.12	$0.38^{\dagger}$	0.32	0.22	0.24	$-0.36^{\dagger}$
	(0.06)	(0.18)	(0.20)	(0.25)	(0.21)	(0.19)	(0.25)
Race: white	0.03	-0.07	-0.17	-0.12	-0.15	$0.30^{\dagger}$	-0.14
	(0.07)	(0.21)	(0.25)	(0.28)	(0.24)	(0.22)	(0.27)
Female	0.11^{*}	-0.89^{***}	$0.41^{\dagger}$	0.03	0.50^{*}	$0.38^{\dagger}$	0.35
	(0.05)	(0.17)	(0.19)	(0.21)	(0.19)	(0.16)	(0.21)
N	1101	1564	1101	1101	1101	1101	1101

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Values are drawn from regression models with acts of participation as the dependent variable. Column two is based on the combined EGSS and OPSP samples, both unweighted; other columns are based solely on the EGSS data. Due to missing data, sample sizes are somewhat smaller than for other analysis.

positive predictor of political blogging, but political efficacy is not. Income is not a significant predictor, but education is. After controlling for other factors, work status, marital status, and race are not significant predictors of whether a person will blog. However, consistent with the recent findings of Schlozman, Verba and Brady (2010), age is a negative predictor of political blogging. Finally, women are significantly less likely to blog than men, even controlling for other factors.

Columns 3 through 7 in Table 2.2 report results from the same model applied to other forms of activism. In this analysis, all the independent variables remain the same, but the dependent variable changes. Therefore, interpreting coefficients as odds ratios, we can compare coefficients across models[22]. Compared to other forms of participation, income and efficacy both have relatively small predictive power for blogging. On the other hand, political interest, education, gender, and age all have unusually large coefficients. The big surprise in this analysis is the substantively small and statistically insignificant coefficient on civic skills—blogging is the only participatory act for which civic skills are not a significant, positive predictor. I will discuss possible interpretations of this result in the next section.

2.7 Discussion

These results in hand, we turn now to implications for political blogging, participation, and communication. Following the directive to develop "theories in the middle range" (Merton, 1968), this section gathers stylized facts and integrates them with other important ideas in the literature on political participation and media. From a methodological perspective, this discussion is the first payoff for all the work done to construct a representative sample of bloggers and survey them in parallel with the U.S. population.

[22] The exception is the intercept term in the blogging model, which is skewed upwards because of our oversample of bloggers. This is a standard feature of case-control designs.

Section 2.7.1 begins by collecting evidence about the social forces that drive political blogging, arguing that—in the broad strokes, at least—they are remarkably similar to those that drive other modes of political activism. From there, Section 2.7.3 pivots to a discussion about socialization and how paths to political participation may differ in the world of online activism. Together, these first two sections explore the notion of blogging as activism: how blogging is similar to, and different from traditional modes of political participation.

Section 2.7.4 develops the theme of quasi-professionalism in the blogosphere, arguing that the blurring of professional boundaries complicates existing theories of political participation and media. Section 2.7.5 concludes by arguing that a similar blurring of boundaries is occurring within political activism and suggesting directions for future research.

2.7.1 Blogging as activism

In a broad sense, the leading finding from this study—the jumping-off point for future theory building—is that for all intents and purposes, political blogging is a form of political activism. Bloggers think like activists; they look like activists; they act like activists. The same constellation of social forces is in play. This is largely a confirmatory result, bringing results from previous work together with new data and systematic analysis.

However, we find that the balance of motivating factors is different in the blogosphere. Relative to most other forms of political participation, blogging is correlated more with political interest and less with political efficacy. This finding suggests that for many bloggers, posting about politics online may be a form of civic sensemaking (Walsh, 2004), rather than a goal-driven form of activism. An alternative possibility is that bloggers hold realistic assumptions about the difficulty of bringing about change in politics, but are interested in participating for intrisic reasons (Li, 2005).

In contrast to most other political acts, income has a negligible effect. Two other acts have similarly small coefficients for income: advertising for political causes, and contacting an elected official. This pattern is consistent with the resource model: after controlling for other variables, acts that make no direct demands on money, and small demands on time are equally accessible to low- and high- income citizens.

From the perspective of representation, this would seem to paint an optimistic picture: blogging is accessible to almost anyone, with little regard for wealth or status. However, as we will discuss in Section 2.7.3, formal education—also an important driver of social stratification—is an unusually strong predictor of participation through blogging. On possibility is that by requiring facility with words but little monetary input, the blogosphere introduces a cross-cutting element into the usual dynamics of socioeconomic status. Another more pessimistic possibility is that the rising generation of online activists is already advantaged by formal education, and will grow into additional earning power as time goes on, leading to even more stratification and representational distortion.

Age also has a negative coefficient, a finding consistent with Schlozman, Verba and Brady (2010). This appears to be the result of a cohort effect among would-be activists. It seems reasonable to suppose that when blogging became a viable political medium, many younger activists adopted it as one of their primary means of political expression. At the same time, some older activists picked up blogging, but the adoption rate among this group was much lower. Very few older adults adopted blogging as a form of participation, even though this group is the most politically active. Given the "young, but not too young" pattern we saw among bloggers earlier, plus the relative novelty of blogging as a medium, this is really the only explanation that fits the available data.

This begs the question whether political blogging (and other online activities) are *replacing* or *augmenting* other modes of participation. On one hand, the "slactivist" hypothesis speculates that among young adults, online participation acts as a low-cost substitute, crowding out more meaningful forms of participation (Morozov, 2012). The opposing viewpoint is that online participation complements and enhances offline participation by enabling rapid dissemination of information and tactics, and providing low-overhead channels for maintaining relationships and mobilizing as the need arises Karpf (2009). The debate between these points of view is far from over, but the evidence to date leans in favor of complementing and enhancing (Ogilvy and Georgetown, 2011). This chapter's finding that bloggers are also avid offline participants lends weight to that position.

2.7.2 Democratic bandwidth

Assuming that blogging is complementing and not substituting for traditional modes of participation, what consequences should we expect? As a high-bandwidth activity[23], political blogs create new channels for input into both elections and the policymaking process itself.

By reading blogs, policymakers can learn about likely policy outcomes and public preferences. By maintaining their own blogs, they can float ideas and get feedback. In addition, the format of blogging encourages linking, "following," and reposting, making it possible for ideas to disseminate through the blogosphere and reach policymakers indirectly.

Focusing on inputs to policymaking provides a nice (partial) refutation to one of the most serious criticisms leveled at political blogging so far. Hindman (2010) has argued that readership among political blogs is low, and even more stratified and unequal than readership among newspapers. These facts are worrisome if we think of blogs solely as a means of *informing the public.* They are much less troublesome if

[23] By *high-bandwidth*, I mean "capable of conveying a complex message." *Channel capacity* in Shannon's information theory offers a precise and conceptually useful technical definition of bandwidth in the context of engineered information transmission (Shannon and Weaver, 1949). Verba, Schlozman and Brady (1995) discuss essentially the same characteristic of participation under the header of "capacity for conveying information."

we think of blogs as means of *informing policymakers.* In the case of policymaking, an audience of one can be sufficient, as long as that one person can influence the final decision. If we allow for non-electoral inputs into public policy, then a fact can influence policymaking without many people knowing it. Overlooking this possibility, Hindman conflates readership with influence.

Readers familiar with the theory of *networked civil society* put forward by Benkler (2006) will recognize several common elements. In Benkler's framework, many-to-many communication invigorates the public sphere, leading to broader intake of ideas, better discussion, and ultimately better governance. Benkler is very critical of the media oligopoly of the mid-20th century, which he says was heavily influenced by money and ideology, and exercised outsized control on public access to information. According to his account, the current proliferation of online information sources is better than being dependent on a handful of corporate broadcasters, even if it still falls short of utopia.

This picture of the public sphere is appealing and not entirely untrue. Unfortunately, Benkler's model ignores the electoral input to public decision making. Benkler treats "government" as a unitary actor, and makes only passing reference to elections and political parties. This is an important omission in a political system where legislators and executives are elected in two-party, zero-sum, partisan elections. Policymakers are influenced not only by the ebb and flow of ideas in public debate, but by their ability win re-election (Mayhew, 2004). We have strong reason to believe that access to additional channels, selective exposure, and ideological pandering are leading to increased polarization in the electorate. (For example, see Prior 2007 and Lawrence, Sides and Farrell 2010). What if this polarizing electoral effect dominates the enriching discursive effect that Benkler outlines?

To summarize, focusing on democratic bandwidth leads us to ask where the information used to make policy decisions comes from. Of particular interest is how society closes the information gap between low-bandwidth elections and high-bandwidth policymaking. As high-bandwidth modes of political participation, blogs and other online media have the potential to provide an important new input into the democratic policymaking process. However, the consequences of this new bandwidth are unclear, because many of the likely effects are cross-cutting and hard to observe directly. In order to understand the normative consequences of this new bandwidth, we need to better understand the ways information interacts with democratic institutions.

2.7.3 New paths to online participation?

Returning to the main argument, most of the differences we've seen are small deviations from the canonical model of political participation. Blogging stretches the resource model in places, but all of these changes seem reasonable given what we know about the format of blogging. One of the centerpieces of the resource model of participation is the notion that people acquire useful civic skills at work, at church, and through other organizations. These non-political institutions have long been incubators for important political skills.

The biggest apparent difference between blogging and older forms of activism is the fact that after controlling for other variables, traditional measures of civic skills do not predict political blogging. This significant finding stands in contrast to the moderate to large statistical effects for every other form of participation measured in these surveys. Best and Krueger (2005) discovered a similar pattern in their 2003 study of online participation. Two explanations are possible: either blogging requires different skills, or bloggers acquire the same set of skills by other means[24].

To the first explanation: the questions asked in the original activist survey focus on giving presentations and speeches, and planning and chairing meetings. In a general sense, these activities may be useful for cultivating interpersonal skills. However, these

[24] I admit the possibility of a third explanation: the insignificant coefficient on civic skills for political blogging is an artifact of the 2SAIV regression estimation strategy. For the reasons given previously, this does not seem the most likely explanation.

activities are not direct prerequisites for political blogging. Perhaps, as Best and Krueger argue, the skills required for online participation are simply different than other forms of activism.

To the second explanation: perhaps other institutions can substitute for work, church, and civic organizations as civic incubators. The "church, work, and other organizations" typology was developed in the early 1990's, before telecommuting, outsourcing, and the Internet began affecting social institutions. It may be that such social technologies have enabled other institutions to acts as socializing paths to political participation. For instances, perhaps online "institutions" such as search engines, online communities, and social networking sites are able to impart skills and interest in politics.

An alternative hypothesis is that formal educational institutions are filling this training and socializing niche for bloggers. Given that bloggers tend to be younger and are far more likely to hold advanced degrees, it is likely that a larger portion of their habits and worldviews have been shaped at universities. The unusually strong relationship between political blogging and education adds some weight to this hypothesis. If so, inculcation of online skills could become a valuable source of "social returns" (Hout, 2012) from formal education. Whether these skills would act as a form of "democratic enlightenment" or a more self-serving form of "democratic engagement" is another question (Nie, 1996).

Reframing political blogging as a form of activism points us in the direction of a host of interesting questions from the normative theory of political participation. In this light, the question of who blogs becomes a question of representation not unlike the question of who votes (Wolfinger and Rosenstone, 1980). Normatively, we would hope that the mix of ideas and interests discussed in the blogosphere would be representative of the community at large (Burns, Schlozman and Verba, 2001). We would also hope that bloggers' voices would be "clear, loud, and equal." (Verba, Schlozman and Brady, 1995).

2.7.4 Blogging as quasi-journalism

The empirical results in this chapter also have strong implications for the old debate about whether blogging is a form of journalism. Based on the evidence we've seen, the "average blogger" is clearly an activist. Traditional norms of objective journalism enforced a barrier between the two roles: with few exceptions, one could be an activist *or* a journalist, but not both at the same time (Schudson, 2001). Therefore, *if* activism and journalism are mutually exclusive activities, then then average blogger is not a journalist. Either way, the best available evidence demonstrates that bloggers are activists first, and journalists second, if at all.

But are activism and journalism mutually exclusive? A few years ago, we might have been content to answer "yes," and close the debate on the nature of blogging. However, these days, the distinction is less clear. Mainstream media channels like Fox News and MSNBC have made ideology an integral part of their business models (Iyengar and Hahn, 2009). Millions of young adults rely on Comedy Central as their primary source of news (Baym, 2005); the same comedians who host these fake news shows sponsor super-PACs and stage rallies in Washington, D.C. From the evidence presented in this chapter, we know that a large percentage of non-professional, ideologically motivated bloggers—activists—create news-like content that attracts significant revenue and readership. These patterns point to a trend of increasing overlap between activism and journalism.

What is causing this trend? A full exploration of all the possible explanations is well beyond the scope of this chapter. For now, let us pursue one line of argument which I will call the *professional-technological hypothesis*.

The basic premise of the professional-technological hypothesis is that journalism, like most other professions, derives prestige and revenue by creating and enforcing boundaries around a given domain of work. Like doctors, lawyers, and accountants, journalists have developed standards for the day-to-day work of professional reporting. These standards are embodied in professional norms, style guides, codes of conduct, etc. and are enforced through institutional control of training, hiring, membership in professional communities, awards, and so on. However, as Zaller (1999) writes, the boundaries of such "professional cartels" are "difficult to maintain, particularly under conditions of rapid social or technological change, and they are doubly hard to maintain in the presence of free market competition."

Again from Zaller:

> Elite journalism is under fire—more-or-less continuous fire—from a mass audience that isn't much interested in politics, lower-status journalists willing to meet the mass audience on its own level, and politicians vying to control their own communication and increasingly adept at doing so. Elite journalists are no patsies in this struggle ... [b]ut they are in a more precarious position than many outsiders realize, and they know it.

These words, written in 1999, have proved prescient. Since then, pressure from cable TV and Internet-based media (including blogging) have dramatically reduced the economic clout and public impact of professional journalists (Prior 2007; Hindman 2010.) This has led to a blurring of professional boundaries, with many quasi-journalists supplying news-like information without adhering to the practices of 20th-century professional reporting (Keen, 2007; Shirky, 2009). Among other changes, many prominent quasi-journalists speak from explicitly partisan or ideological viewpoints, combining reporting and "journalism" with advocacy and activism (Karpf, 2009). This state of affairs that could not have existed under the old rules of journalism.

So, are bloggers journalists? Unlike activism, journalism has been defined in terms of professional boundaries. As these boundaries blur and evolve, the semantics of the question become difficult. One reasonable answer is that bloggers are not journalists according to any traditional definition—but the same is true of many prominent voices that provide news in the modern American media ecosystem.

If we accept the professional-technological hypothesis, then blogging is best seen as part of a growing movement of technologically enabled quasi-journalism that competes with traditional journalism for revenue and influence. Which, if any, forms of quasi-journalism will be included within future definitions of professional journalism remains an open question.

2.7.5 Research implications of the professional-technological hypothesis

As we segue into the conclusion, it is worth developing the professional-technological hypothesis in one additional direction: blurred professional boundaries for political activists. Political activists have never institutionalized professional boundaries in the manner of journalists. Nevertheless, activism in the late 20th-century was marked by a reasonably clear divide between professionals and volunteers (Bimber, 2003)[25].

This conceptual division between volunteers and professionals has provided a convenient theoretical and methodological line for researchers. Within the field of public opinion, the professional boundary is demarcated by the theoretical distinction between "political elites" and "the mass public" (e.g. Stimson (2004); Zaller (1999)). Interestingly, research methods for studying each group have tended to diverge: interviews, ethnographies, and close reading of primary documents for elites, in contrast to surveys and experiments for the mass public. Both approaches have produced useful results, but in terms of volume of research and citations, it seems fair to say that surveys and experiments are the dominant methodology in public opinion. Consequently, most of the theory of public opinion explain nuances of behavior in the mass public, with considerably less attention to the behavior of political elites.

A similar divide is evident in the field of political participation. Once again, re-

[25] As Bimber notes, this divide was probably the product of many of the same mass-media technologies that led to the rise of professional journalism.

searchers have drawn a conceptual line at the boundary between professional activists and volunteers, and again, surveys of nonprofessional activists have been the primary research methodology. A brief review of theory-building studies and political participation literature illustrates this pattern quite forcefully: Wolfinger and Rosenstone (1980); Verba and Nie (1987); Rosenstone and Hansen (1993); Verba, Schlozman and Brady (1995); Burns, Schlozman and Verba (2001). Each of these studies is based on survey methods, and close methodological reading reveals that each is primarily focused on nonprofessional activists.

This common theoretical and methodological underpinning deserves attention because it presents both a threat and an opportunity to this literature. The threat arises from the fact that virtually the whole field hangs on a common assumption: we can draw a clean distinction between professional and volunteer activists. Blurred boundaries around "elite" or "professional" activism would challenge this core assumption. A priori, it is not clear whether and to what extent theories of activism developed in the context of nonprofessional participation will explain the behavior of quasi-professionals.

Our findings on political blogging demonstrate that the lines are blurring. What should we make of a political blogger who earns several thousand dollars per year, while working full-time in an unrelated position? Or a blogger who devotes dozens of hours a month to following and reporting on city politics? Beyond blogging, other forms of online activism show a similar uptick in participation. For example, the 2012 Presidential campaign shattered records for online donations, volunteer hours, and events organized online (Scherer, 2012). Beyond huge numbers of small participatory acts, the campaigns worked actively to engage supporters repeatedly, transforming donors to canvassers to volunteer organizers.

Such quasi-professional activists have always existed; anecdotal evidence suggests that they are becoming increasingly common. As a percentage of the population, quasi-professional activists probably remain a small minority, but it seems reasonable to suppose that they exert outsized influence on campaigns, media, and policymaking. Moreover, as a group on the border between political elites and the mass public, activists are likely to be disproportionately affected by these trends.

In order to address this threat to the validity, theories of political participation will need to consider professional incentives and socialization more carefully. This chapter has taken a first step in that direction by examining revenue, attention, and popularity among political bloggers. Exploring these incentives and their effect of bloggers'—and other activists' patterns of behavior—strikes me as an important avenue for further research. In a different vein, our discoveries about the unusually high predictive power of education and unusually low predictive power of civic skills for explaining participation in the blogosphere are also suggestive of the types of training and socialization necessary to be active in the blogosphere. One promising line for future research would be to explore skill transfer from educational settings to new forms of political participation, in the spirit of Schlozman, Burns and Verba (1999).

2.7.6 Future directions

In closing, let me highlight on the opportunity that strikes me the single most promising direction for future research. Given that political bloggers fit the profile of political activists, it seems reasonable to suppose that we can use behavior observed in the blogosphere to extend theories of participation. As noted above, the vast majority of research on political participation focuses on *levels of participation*: Who shows up and how much do they contribute? Very little research attends to the *content of participation*: What do people actually say and do once they arrive on the public stage?

In future research, content analysis could investigate aspects of participation such as civility (Sanders 1997; Carpini, Cook and Jacobs 2004; Sobieraj and Berry 2011), selective exposure (Frey 1986; Lawrence, Sides and Farrell 2010), adherence to norms of traditional journalism (Schudson 2001; Gillmor 2006), and attention to different topics over time (Lippmann 1927; Iyengar and Kinder 2010; Leskovec, Backstrom and Kleinberg 2009). These aspects of participation matter deeply for democracy, but have been difficult to study in other contexts. Success in this line of research could combine the best of Walsh's ethnographic studies of informal political discussion (Walsh, 2004) with Mutz' detailed investigation of cross-cutting political exposure (Mutz, 2006). It could measure directly the forces of "elite discourse" that have typically been inferred indirectly from time series or panel data (e.g. Zaller 1992; Stimson 2004; Prior 2007, Page and Shapiro 2010).

Because the political blogosphere offers a wealth of easily validated, high-bandwidth data, it provides a natural laboratory for studying the content of participation. This laboratory would enable us to take many aspects of the social system surrounding public opinion and political participation and observe, measure, and theorize about them directly.

2.8 Conclusion

This chapter establishes a crucial link between political bloggers and political activists. Multiple lines of evidence demonstrate that bloggers think like activists, look like activists, and act like activists. These results pave the way for new directions in the study of political participation: the content of participation, bandwidth, and the role of information in democracy.

CHAPTER III

Methods for very large-scale content analysis

3.1 Introduction

By its nature, content analysis presents special difficulties for scientifically valid measurement. The process is inherently subjective, and must therefore be based on the judgments of human coders who make mistakes, get bored, try to game the system, and often have very different perceptions. All of these human factors complicate the process of deriving valid measures from text. In this project, we have the additional requirement of applying those measures to millions of blog posts.

This chapter details methods for meeting the twin challenges of validity and scale in content analysis. The main story in this chapter is about reliability and accuracy. I describe the process of expert, novice, quorum, and automated coding, and demonstrate that most of our codebook items perform quite well in terms of intercoder reliability. Along the way, I take short detours to discuss practical issues related to gathering data and assessing reliability: data structures, identifying spammers, and specific items with poor reliability scores.

3.2 Research questions

This chapter describes (1) a series of codebooks for journalistic writing and civility in blog text, (2) tests to ensure that those cookbooks yield reliable measures, and (3) methods for scaling those measurements to millions of blog posts. In this endeavor we have three primary research questions:

- Can expert coders reliably measure newswriting and civility in the text of blog posts?
- Can novice coders with no prior exposure to the theory of constructive discourse and journalism also reliably measure the same variables?
- Can automated text classification replicate expert and novice coding with high accuracy and low bias?

To ensure that our measures are both valid and scalable, each of these steps must be completed successfully. Reliable coding among experts is necessary for primary codebook development. Reliable coding among novices is necessary to guarantee that the methodology satisfies the scientific criteria of transparency and replicability. Accurate and reliable coding by automated classifiers is necessary in order to apply the measurements to a corpus of millions of blog posts.

3.3 Methodology

3.3.1 Expert codebook development

The content of political blog posts can be evaluated along countless dimensions. For purposes of this dissertation, I choose to focus on two dimensions in particular: newswriting and civility. These concepts occupy important positions in many theoretical debates about political discourse and new media. For our purposes in this chapter,

Table 3.1: Items for the four newswriting codebooks

Neutral voice
Does the author use the first person (e.g. I, we, our) outside of quotes?
Does the author express his/her opinion directly in this text?
Public import
This article is about a topic that affects many people.
The article is about an issue of government policy.
This article talks about what government should or should not do.
This article is about a political issue.
This article is on a lifestyle topic, such as fashion, entertainment, food, etc.
This article focuses on offering helpful advice to readers.
Information gathering
Outside facts, sources, and evidence are an important part of this article.
Personal experiences from the author's own life are an important part of this article.
The author appears to have spent significant effort gathering information for this article.
This article is based on information that is not available on the Internet, such as interviews and on-the-ground reporting.
Sources and evidence
Direct quotations from people other than the author
Personal experiences from other people's lives
Statements by experts
Statements by eyewitnesses, or people who have been directly affected by the topics discussed in the article.
Quantitative information, such as percentages, prices, poll results, etc.

they will serve as a diverse set of measures to test new methods for large-scale content analysis.

By *newswriting*, I mean "patterns of writing that reflect the norms, traditions, and professional standards of mainstream American 20th century journalism" (Schudson, 1981). Newswriting in blogging is important because the news industry is changing rapidly, with blogging encroaching on a media niche that used to be occupied by professional journalists (Meyer, 2009) (Hindman, 2010). Therefore, the long-term impact of blogging is closely tied to the ways in which bloggers choose to follow or defy the tradtional conventions of newswriting (Karpf, 2008).

Table 3.1 lists items for each of the four newswriting codebooks: neutral voice, public import, information gathering, and sources and evidence.

Civility is harder to define, in large part because the boundaries and dimensions of the concept are contested in deliberative theory (Sanders, 1997) (Mutz, 2008). For our purposes, I focus on three relatively well-defined aspects of civility: divisiveness, respect for others, and appeals to anger and fear[1]. These concepts are closely related to the normative ideals of inclusion, veracity, transparency, and rationality in public discourse. Although various schools of thought debate the relative importance of these aspects of civility, they all agree that certain patterns of communication are essential for opinion formation and healthy democratic self government.

Codebook development was an iterative process, centered on a group of "expert coders." Throughout this chapter, I will use the term "expert coder" to refer to those involved in creating the codebooks: myself, plus undergraduate research assistants. This research team went through multiple rounds of testing and evaluation for each of the seven codebooks. In each round, we sought to revise codebooks to (1) capture the core concepts of theories of constructive discourse, and (2) enable high-reliability content coding within the team.

Once codebook development was complete, I evaluated intercoder reliability on the final versions of each codebook as follows. First, I drew random samples of 60 previously unseen blog posts containing at least 1,500 characters (approx. 250 words)[2]. Next, for each codebook, a pair of expert coders independently coded each blog post in the sample. Finally, we measured intercoder reliability using the Krippendorff's alpha statistic.

[1] In addition to these codebooks, I also attempted to create codebooks for other aspects of civility, including reciprocity (Gutmann and Thompson, 2009) and violent rhetoric (Kalmoe, 2011). However, these proved extremely difficult to measure reliably in blog content, even among expert coders.

[2] Our early experiments in codebook development showed that very short blog posts don't contain enough content for either human or automated coders to produce reliable labels. In addition, these very short posts tend to be of less substantive interest.

Table 3.2: Items for the three civility codebooks

Divisiveness
The author takes a clear stance on the issue.
To what extent does the author frame the issue as an “us against them” situation?
On the issues discussed in the article, the author seems open-minded.
This author blames specific people or groups for bad outcomes.
The author sounds like s/he would be open to compromise on this issue.
The author of this article seems like a reasonable person—someone you could have a good discussion with, even if you disagreed
Respect for others
This text mentions opposing viewpoints.
The author shows respect for people holding opinions that oppose his/her own.
The author shows respect for opposing viewpoints.
Much of this text is spent criticizing others’ motives (e.g. “he’s out to get us,” “she can’t be trusted,” “greedy,” “dishonest,” etc.)
Much of this text is spent criticizing others’ competence (e.g. “he’s an idiot,” “she is clueless.”)
The author uses derisive labels for people.
Appeals to anger and fear
The author tries to make the reader feel angry about this issue.
The author uses emotionally-charged language.
The author talks about threats to cherished values (e.g. political, religious, moral ideals).
The author talks about threats to physical well-being and safety.
The author talks about threats to personal economic interests (e.g. jobs, income, tax rates, etc.).
The author uses exaggeration and/or hyperbole (“the worst idea ever,” “everyone is talking about it.”)

3.3.2 Novice content analysis

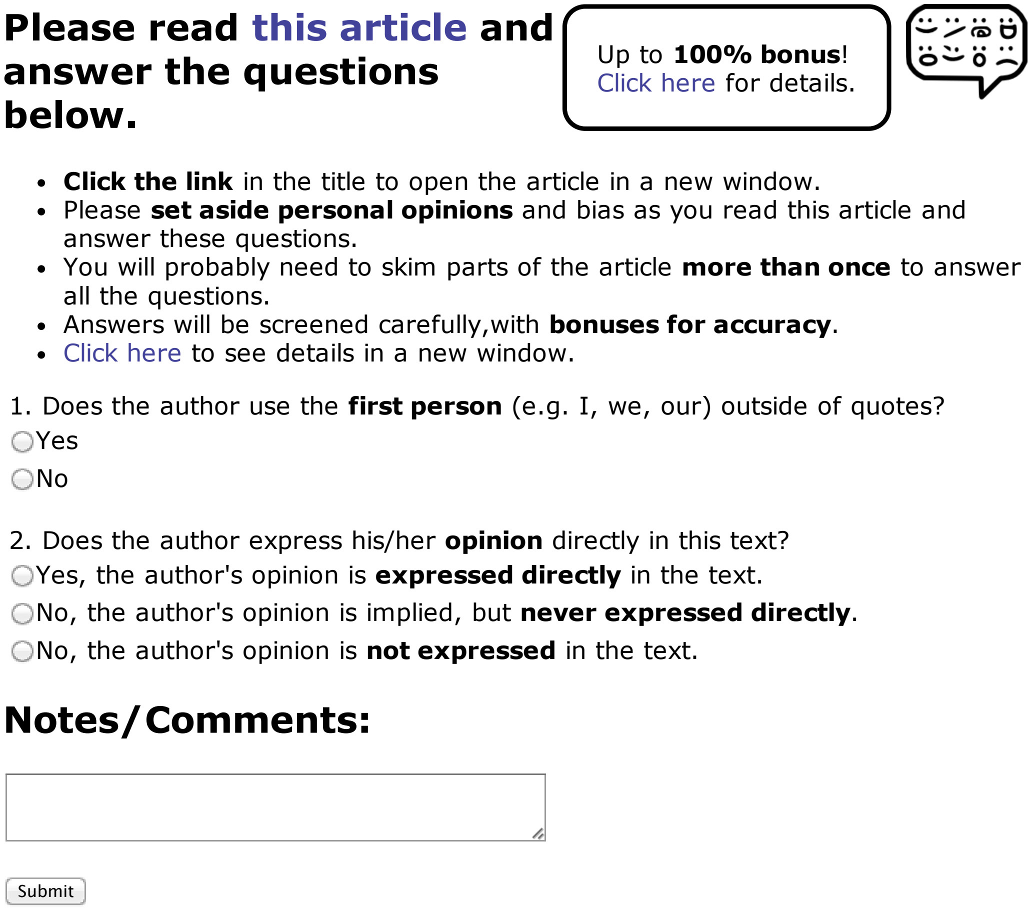
The advantage of expert coding is that it is relatively easy to develop a common language, achieve consensus, and label documents with high reliability. However, the process isn't fully replicable, and it doesn't scale well—it's one thing to organize research assistants to code dozens or hundreds of articles, and something else to persuade them to code thousands or tens of thousands. To scale up the volume of coding by two orders of magnitude, I took the same codebooks to the crowdsourcing site Amazon Mechanical Turk (MTurk), and recruited workers to code a comparable set of blog posts. I will refer to these workers as "novice coders" or "turkers."

The decision to use Mechanical Turk as a platform for content analysis was driven by two primary considerations. First, MTurk provides a practical means of ensuring replicability. Workers on MTurk have no prior exposure to my codebooks or the objectives of the study. Employing them is an effective way of "double-blinding" the content analysis portion of study against subconscious bias. A second advantage of MTurk is its cost-effectiveness: piece rates on MTurk are low enough that I was able to recruit dozens of coders to assign labels for thousands of blog posts. For each of the seven codebooks, a corpus of 2,380 documents was labled on MTurk. As described in Section 3.3.4, many of these documents were labeled by multiple coders.

The downside of MTurk is that intercoder reliability is typically much lower, especially on nuanced and subjective tasks. Accordingly, I developed an MTurk task interface designed to be user-friendly, intuitive, and clear, in order to minimize distractions, confusion, and ambiguity that could lower intercoder reliability. Figure 3.1 shows a screenshot of the MTurk task for the neutral voice codebook. Other codebooks followed the same format; screenshots can be found in Appendix D.

To further improve reliability, I restricted the task to turkers who had completed at least 100 tasks with an acceptance rate of 90 percent. Futhermore, only turkers

Figure 3.1: Screen shot from the neutral voice codebook



Please read this article and answer the questions below.

Up to **100% bonus**! Click here for details.

- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

1. Does the author use the **first person** (e.g. I, we, our) outside of quotes?

○Yes

○No

2. Does the author express his/her **opinion** directly in this text?

○Yes, the author's opinion is **expressed directly** in the text.

○No, the author's opinion is implied, but **never expressed directly**.

○No, the author's opinion is **not expressed** in the text.

Notes/Comments:

with accounts in the United States were allowed to participate[3]. For the first five codebooks, I paid \$0.06 per task with the promise of "up to 100% bonuses" based on each coders' quality and quantity of work. On the remaining two codebooks, I raised the rate of base pay to \$0.10. Additional details were supplied in a FAQ (see Appendix D). In the end, turkers earned an average of \$0.1124 per task, with an average base rate of \$0.0726 and an average bonus rate of \$0.0397.

In addition, I took steps to remove spammers from my population of workers, and to combine worker scores using a technique I call "quorum coding." Together, these techniques are sufficient to bring novice intercoder reliability scores up to levels comparable with expert coding. I will discuss the details of these methods in the results section of this chapter.

3.3.3 Automated coding

By reporting intercoder reliability and conducting our content coding with novice coders, rather than experts, we have already established a higher bar for replicability than the vast majority of published content analysis studies. However, we're not done yet, because we have "only" coded a few tens of thousands of blog posts. Ultimately, the research design calls for reliable coding of millions of blog posts. The last step is to train text classifiers that mimic the novice coders. If we can build accurate enough text classifiers, then there's nothing stopping us from applying them to as many millions of blog posts as we want.

Accordingly, I trained a separate classifier for each codebook item. Each classifier was based on a LASSO regression model[4] with a regularization constant of .015[5]. The

[3] Other researchers have found that turkers outside the U.S. often produce less reliable results (Shaw, Horton and Chen, 2011). This finding agrees with my own experience as well.

[4] As a member of the regression model family, LASSO models expect the dependent variable to be a continuous, interval-level score. Since our data are technically ordinal, an ordinal classifier would be a better statistical fit. However, training such models is computationally prohibitive. In practice, treating ordinal scales as interval data seems to work well enough.

[5] Classifier accuracy could probably be improved slightly by tuning this parameter for each model.

feature set includes 3,112 characters, character bigrams, and alphanumeric words that occurred in at least 100 documents within the training and testing corpus. I mapped the text to lowercase, but did not apply any other stemming or stopping to the feature set[6]. Following common practice, I used binary presence vectors ("Did the feature occur or not?") rather than count vectors ("How many times did the feature occur?").

After each classifier was trained, I applied linear recalibration to correct for overfitting—or possible underfitting due to the regularization constant—as follows[7]. I first fit a bivariate linear model regressing predicted values onto actual labels within the *testing* set[8]. This model yields two parameters, a slope and an intercept, and can be applied to all classified values to ensure that they are measured on the same scale as the original labels assigned by human coders. By construction, rescaling does not affect the correlation between human and automated coding.

This fact is important because it provides a convenient way for us to measure the reliability of classifiers[9]. It happens that the Krippendorff's alpha and Pearson's R statistics are exactly equal, in the special case of continuous measures with labels assigned by two coders with the same means and standard deviations, and no missing values. After recalibration, our measures satisfy all of these criteria. Therefore, we can use the Pearson's R statistic as a measure of classifier accuracy. The statistic is directly comparable to the Krippendorff's alpha scores we use to evaluate intercoder reliability among expert and novice coders. This unexpected mathematical equivalence is a useful byproduct of this interdisciplinary statistical investigation.

[6] On the whole, my goal was to achieve reasonably high classifier accuracy, without perfectly honing each individual text classifier.

[7] This technique is very similar to the one described in Chapter I. The only difference there is a minor adaptation required to suit a classifier to a binary variable rather than a continuous variable.

[8] Applying this regression to the testing data does not introduce any significant risk of overfitting, because the number of parameters (two) is very small relative to the number of observations (typically, several hundred).

[9] When discussing automated text classifiers, I will use the terms "reliability" and "accuracy" interchangeably, to refer to the classifiers' ability to match novice coding.

3.3.4 Data structures for scale and reliability

Before proceeding with the rest of this chapter, is worth describing in detail the data structures which I used for novice coding. Like most content analysis projects, this research design presents two competing criteria for data collection. First, we need to assess intercoder reliability—ideally, reliability for each coder individually. Second, we want to label as many documents as possible for further statistical analysis (e.g. trainining text classifiers). The first criterion implies that we should code each document as many times as possible. The second says we should code each document only once. Given a limited budget for content coding, how many times should we code each document?

To meet both of these needs, I developed a structure that I call "5-by-25 replication." Specifically, 25% of the blog posts were coded 5 times, and 75% were coded once. Assignment to the five-coded or once-coded condition was random, and novice coders did not know whether any given document was once- or five-coded. In terms of cost, this approach is equivalent to coding each document twice, but this structure gives us additional analytical leverage on our key problems. The five-coded documents allow us to assess intercoder reliability, and the once-coded documents give us plenty of training cases for the classifiers[10].

I find that 5-by-25 replication is a much more effective data structure than traditional 2-by-100 replication, for three reasons. First, intercoder reliability is typically calculated using pairwise comparisons among coders. When two coders code a single document, we gain one coding pair (coder A vs coder B). When five coders code a

[10] For other content analysis projects, other similar approaches (e.g. 4-by-33, 10-by-20) might be appropriate. I have made no attempt to identify the optimal balance for replicated coding. Instead, my goal here is to show that for *this* research project, 5-by-25 dominates traditional 2-by-100 replication.
With that said, I strongly suspect that this problem is subject to optimization, given a modest amount of information about coder and document characteristics. Lamberson and Page (2012) provide analysis of optimal coder accuracy and diversity relative to sample size. This approach, combined with the estimation techniques of Welinder et al. (2010), seems like a good jumping off point.

single document, we gain 10 coding pairs (A vs B, A vs C, B vs C, A vs D, etc...) On average, coding each document twice generates one code pair per document, and 5-by-25 coding generates 2.5 code pairs per document—a 150% improvement at no additional cost.

Second, if we decide that a given coder's work should be removed from the data set, 5-by-25 coding is much more resilient than pairwise coding. For documents coded five times, losing one coder removes 40% of our coding pairs. For documents only coded twice, losing one coder leaves us with a single score and no pairs at all.

Third, 5-by-25 replication enables us to average document coding across quorums of five coders, rather than quorums of two. As we will see, this allows us to construct testing sets with higher underlying reliability.

There is one minor drawback to 5-by-25 replication: individual coders may need to code more documents before we can accurately assess their reliability. If only one in four documents allows for assessments of reliability, and we need a sample of k (e.g. ~ 20) documents to judge individual coders, then each coder must code at least $4k$ documents. In a come-as-you-please setting like Mechanical Turk, coders might quit before finishing the requisite $4k$ tasks, leaving us unable to estimate coder-specific reliability. This could leave a serious gap in our chain of evidence, and cast doubt on the validity of our measurements.

Fortunately, another characteristic of Mechanical Turk mitigates this problem: the workload distribution on MTurk is strongly skewed, with a small number of turkers completing far more assignments than the others. Since these turkers complete so many tasks, we have no trouble assessing their reliability. And since most of our coding is completed by these high-volume turkers, most of our coding is covered by these reliability checks.

In the end, 5-by-25 replication is a cost-effective means of balancing the competing needs of scale and reliability assessment. It introduces substantial new efficiencies with only minor drawbacks, allowing us to stretch our research dollars much further than they would otherwise go.

3.3.5 Errata

To close the methodology section, I highlight two aspects of data collection that will be important later in the chapter.

First, experts and novices coded different sets of documents. Expert documents are a sample from blogs parsed early in the data collection effort. The novice documents are a random sample of all blog posts. This data decision was made so that expert codebook development could proceed in parallel with blog parsing. As a result, the two sets are similar enough to enable comparisons of levels of intercoder reliability, but direct comparisons of novice and expert coding are not possible at the document level.

Second, in addition to blog posts, experts and novices both coded articles from a variety of non-blog sources drawn from Lexis-Nexus. These articles play a small role in this research project, but might provide a useful reference set for future research. Reliability scores for most codebook items were comparable for Lexis-Nexis articles and blog posts. On average, expert and novice intercoder scores were slightly higher for Lexis documents than blog posts.

3.4 Results

3.4.1 Reliability among expert and novice coders

The first column of Table 3.3 shows expert intercoder reliability scores for all codebooks items. For most of the codebook items, expert intercoder reliability was quite high: in the .6 to .8 range, with a handful of lower reliability items. These results are encouraging, because they suggest that many of the important theoretical constructs underlying newswriting and constructive discourse are amenable to measurement.

Unfortunately, intercoder reliability on MTurk is much lower, in the .2 to .4 range, even after I removed coders who were clearly not bothering to read the articles. Columns two and three of Table 3.3 report intercoder reliability scores for codebook items, assessed before and after excluding very low-reliability turkers, as described in the next section.

3.4.2 Coder-specific reliability and filtering

Spamming (or *scamming*) is an important concern for crowdsourced data collection. On MTurk and similar sites, it is not uncommon for workers to skim quickly through tasks, producing a high volume of very low-quality work. From the perspective of the workers, spamming can be a "rational" way to improve hourly income. Some spammers choose answers purely at random; others skim directions and content and make rapid guesses. In either case, the high volume and poor quality of work produced by spammers can cripple a crowdsourced research project. In order to reduce this threat to the validity of our research design, we must be able to identify and filter out spammers.

Fortunately, the same data structures and statistics that we used to measure intercoder reliabilty lend themselves nicely to this task. To identify spammers, we need to measure two variables at the level of individual coders: volume of coding ("How many documents did this coder complete?") and reliability. Volume is easy to measure using the administrative records generated by MTurk. To estimate reliability, I calculated the pairwise Krippendorff's alpha for each coder. That is, I applied the normal Krippendorff's alpha calculation to each coder's labels and all other labels on the same set of documents. This is identical to the normal Krippendorff's alpha statistic, except that each pairwise comparison involves the coder in question.

Table 3.3: Intercoder reliability for codebook items by coding type

Item	**Expert**	**Novice**	**No spam**	**Quorum**	**Auto.**
First person	.781	.671	.699	.921	.634
Neutral voice	.713	.519	.547	.858	.685
Affects many people	.440	.150	.330	.712	.405
Government policy	.814	.165	.471	.817	.520
What gov't should do	.709	.098	.402	.771	.514
Political issue	.777	.476	.543	.856	.617
Lifestyle topice	.657	.106	.400	.769	.535
Helpful advice	.534	.066	.154	.477	.305
Scope	.646	.412	.474	.819	.466
Outside facts	.660	.098	.120	.406	.349
Personal experiences	.755	.288	.305	.687	.551
Effort gathering	.894	.272	.267	.646	.674
Non-Internet info.	.615	.042	.047	.197	.261
Direct quotes	.908	.378	.378	.752	.637
Others' experiences	.804	.249	.266	.644	.385
Expert statements	.719	.059	.089	.328	.302
Eyewitness statements	.875	.149	.224	.591	.296
Quantitative info.	.769	.500	.517	.843	.586
Clear stance	.619	.128	.136	.441	.198
Us versus them	.790	.363	.363	.740	.335
Open-minded	.785	.083	.100	.358	.091
Assigns blame	.759	.326	.332	.713	.381
Open to compromise	.835	.041	.052	.215	.043
Reasonable person	.783	.078	.090	.331	.000
Opposing viewpoints	.551	.214	.378	.753	.540
Respect for people	.680	.009	.221	.587	.094
Respect for viewpoints	.667	-.007	.255	.631	.107
Criticizes motives	.361	.227	.428	.789	.407
Criticizes competence	.176	.163	.316	.698	.251
Derisive labels	.645	.128	.348	.728	.130
Feel angry	.825	.190	.389	.761	.438
Emotional language	.680	.206	.324	.706	.457
Values threat	.849	.227	.347	.727	.477
Physical threat	.887	.225	.310	.692	.335
Economic threat	.701	.192	.338	.719	.590
Exaggeration	.420	.074	.145	.459	.249
Mean (All items)	**.697**	**.210**	**.309**	**.642**	**.385**

Figure 3.2 shows reliability and coding volume for each coder on the neutral voice codebook. This reliability-versus-volume scatterplot follows a typical pattern. Reliability scores are heteroskedastic with respect to volume: the more documents a coder has completed, the more precise our estimate of his/her intercoder reliability. Therefore, among coders with many completed tasks, reliability scores converge to a narrower range. However, there still exist substantial differences among coders' reliability scores. Because these differences are estimated across hundreds or thousands of comparisons with other coders, we can be confident that they reflect real differences in the quality of individuals' coding.

When a coder produces a high of volume coding with substantially lower accuracy than normal, we can deduce that he/she is a spammer. In Figure 3.2, the two coders in the mid-left section of the graph are spammers. These are turkers who completed a moderately large volume of coding (around 50 tasks each) with much lower reliability than average[11].

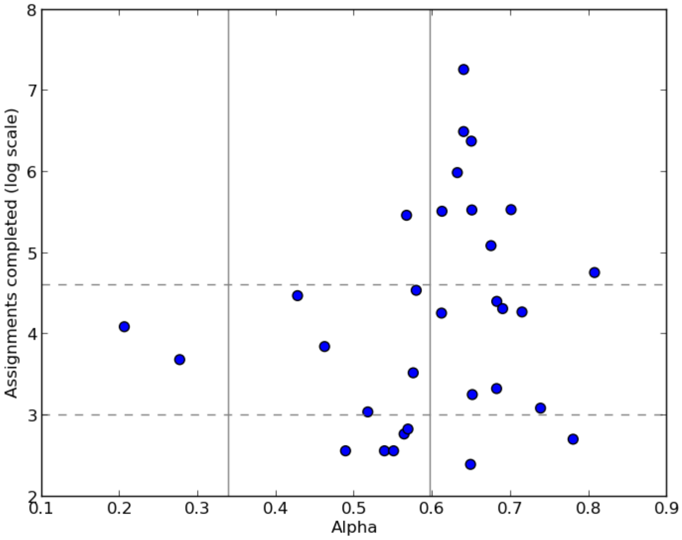
I used two criteria to filter spammers out of my coding pool: statistical significance and substantive significance. In order to be labeled a spammer, a turker had to fail both tests. I will describe both tests in turn.

First, statistical significance: "Can we reject the null hypothesis that the coder's reliability score is equal to or greater than the mean?" This first criteria is important because we want to be confident that a worker is spamming before rejecting his/her work. Also, dropping any turker with below-average intercoder reliability would artificially inflate our estimates of reliability. (For this reason, I've reported intercoder

[11] I strongly suspect that this pattern of moderately high volumes of very low quality work is part of a calculated approach to scamming on MTurk. On MTurk, requesters have the option to reject or accept any task. Rejected tasks are not paid for, and hurt the turker's accept/reject ratio, which can make it harder to find work in the future. However, few requesters actually reject HITs, because of the practical difficulty separating honest turkers from spammers.

Given this incentive structure, it makes sense for spammers to spread their work across tasks from many requesters. They focus enough to get fast at specific task types, but don't invest so much time in any single task that rejection by the requester would severely hurt their income or accept/reject ratio.

Figure 3.2: Scatter plot: Volume and reliability by reliability, for novice coders of the neutral voice codebook



The dashed horizontal lines denote 20 and 100 posts coded, respectively. The solid vertical bars denote the mean score for the group, and mean score minus two standard deviations. The two coders in the middle-right section of the plot are scammers.

reliability scores both with and without spammers.) I used a p-value of .1—low enough to reduce the likelihood of false positives, but high enough to catch most spammers.

Second, I filtered based on substantive significance: "Was the coder's reliability score at least two standard deviations worse than the mean?" The second criteria is important because we want to distinguish between the coders who are merely mediocre, versus those who are truly unreliable.

Univariate outlier detection is a notoriously difficult problem. However, for this application, we have three reasons to believe that this solution worked reasonably well. First and most importantly, most of the turkers identified as spammers had reliability scores that were two or three standard deviations below from the mean. These coders weren't just worse than average, they were *a lot* worse.

Second, reliability scores skew downward. There were several high-volume coders with reliability scores far below average, but none with reliability scores far above average. This is consistent with spamming as a mechanism for unreliability: coding by spammers generates reliability well below average, but there is no analogous source of coding with reliability well above average.

Finally, a piece of anecdotal evidence: not one turker complained when I labeled their work as "unreliable." In the past, I've found turkers to be a vociferous lot, angrily firing complaints and emails back if they feel that their work has been unjustly rejected. The fact that I sent warning emails and rejections to so many turkers without getting a single reply suggests that these reliability checks did a good job identifying spammers.

3.4.3 Quorum coding

As we saw earlier, the intercoder reliability scores from novice coding scores are quite poor. Removing spammers from the mix improves reliability scores, but not enough to be fully confident in all our measures. Fortunately, 5-by-25 replication and the law of averages gives us another avenue for improvement. Consider the following thought experiment: "If several groups of k novices each coded documents, and then we averaged the codes within each group, what level of intercoder reliability would we expect among the group averages?" When we average across groups, it turns out that even small groups have much higher intercoder reliability than individuals. This is the familiar logic of sample sizes, power analysis, and confidence intervals from survey and experimental research. Similar approaches have been used successfully in the field of expert forecasting (Clemen, 1989) and more recently in the emerging field of crowdsourced data collection (e.g Sheng, Provost and Ipeirotis, 2008, Welinder et al., 2010, Karger, Oh and Shah, 2011). With properly structured content coding, we can apply the same logic to content analysis.

Let us introduce some formal notation in order to clarify this logic and add precision to the argument. Let $x_i^{H_j}$ denote the label for variable x assigned to document i by coder j. We model $x_i^{H_j}$ as the sum of two random components: x_i and ϵ_{ij}, the true value of the document with respect to x, plus a coder-specific error for that document.

$$x_i^{H_j} = x_i + \epsilon_{ij} \tag{3.1}$$

Let us use X^H as shorthand for a process of generic novice coding: "random coders from our pool of novices assign labels for codebook item X to a random set of documents from our corpus." Here we are less concerned with the individual document labels, and more concerned with the properties of novice coding as a generic process. We assume that coders are assigned to documents at random, and that the number of documents coded by any one coder is small relative to the total number of documents. If we replicate this process with the same documents and new coders, we can use subscripts to denote the replication: X^{H_1}, X^{H_2}, etc. To extend the notation, let X^{kH} denote a process on generic novice *quorum coding*: "k random coders from our pool of novices assign labels for codebook item X to a random set of documents from our corpus, and we average these labels to get a *quorum label.*" Quorum coding is identical to generic novice coding, with the added step of averaging within document scores across multiple coders.

So far this is all notation—an almost fully generic model of coding. The only substantive assumptions that we've made are that a large pool of novice coders exists, documents are assigned to coders at random, and coder labels are measured such that they can be averaged together into a quorum label.

Let us introduce one further assumption: x and ϵ are drawn independently from separate distributions was finite means and finite, non-zero standard deviations. The means will not concern us much[12]; we will refer to the standard deviations as σ and τ, respectively. In the context of MTurk, these independence assumptions seem

[12] For certain applications, unbiased errors (i.e. $E[\epsilon] = 0$) will be important. However, we can draw a surprising number of conclusions even from data collected from biased coders. We will take up this question again in our discussion of subjective novice coding in section 3.5.2.

quite reasonable. Turkers never see each others' labels, and the process for assigning documents to turkers is largely random, outside the turkers' control.

With this formal framework in place, we can derive some fairly strong conclusions about intercoder reliability. Let $R(X^{H_1}, X^{H_2})$ denote the correlation between X^{H_1} and X^{H_2}: the intercoder reliability of novice coding. Then, $R(X^{H_1}, X^{H_2})$ can be expressed quite simply in terms of σ and τ.

$$\begin{aligned} R(X^{H_1}, X^{H_2}) =& cov(X^{H_1}, X^{H_2})/var(X^{H_1})var(X^{H_2}) && (3.2)\\ =& cov(x+\epsilon_1, x+\epsilon_2)/\sqrt{(\sigma+\tau)(\sigma+\tau)} && (3.3)\\ =& cov(x,x)/(\sigma+\tau) && (3.4)\\ =& \sigma/(\sigma+\tau) && (3.5) \end{aligned}$$

Similarly, the intercoder reliability of a novice quorum of size k ($R(X^{kH_1}, X^{kH_2})$) can be expressed as a function of σ, τ, and k.

$$\begin{aligned} R(X^{kH_1}, X^{kH_2}) =& cov(X^{kH_1}, X^{kH_2})/var(X^{kH_1})var(X^{kH_2}) && (3.6)\\ =& cov(x+\epsilon_1/k, x+\epsilon_2/k)/\sqrt{var(x+\epsilon_1/k)var(x+\epsilon_2/k)} && (3.7)\\ =& cov(x,x)/\sqrt{(\sigma+\tau/k)(\sigma+\tau/k)} && (3.8)\\ =& \sigma/(\sigma+\tau/k) && (3.9) \end{aligned}$$

With both of these formulae in place, we can derive an expression for $R(X^{kH_1}, X^{kH_2})$ in terms of $R(X^{H_1}, X^{H_2})$ and k. In other words, $R(X^{H_1}, X^{H_2})$ is a sufficient statistic for $R(X^{kH_1}, X^{kH_2})$ for any k: if we know the intercoder reliability for individual novice coders, we can project the intercoder reliability of novice quorums of any size. (For this derivation, we will abbreviate $R(X^{kH_1}, X^{kH_2})$ as R_k.)

$$R_k = \sigma / (\sigma + \tau / k) \tag{3.10}$$

$$R_k(\sigma + \tau / k)/(\sigma + \tau) = \sigma / (\sigma + \tau) \tag{3.11}$$

$$R_k(k\sigma + \tau)/(\sigma + \tau) = k\sigma / (\sigma + \tau) \tag{3.12}$$

$$R_k \left[(k-1)\sigma/(\sigma + \tau) + (\sigma + \tau)/(\sigma + \tau)\right] = kR_1 \tag{3.13}$$

$$R_k \left[(k-1)R_1 + 1\right] = kR_1 \tag{3.14}$$

$$R_k = kR_1 / \left[(k-1)R_1 + 1\right] \tag{3.15}$$

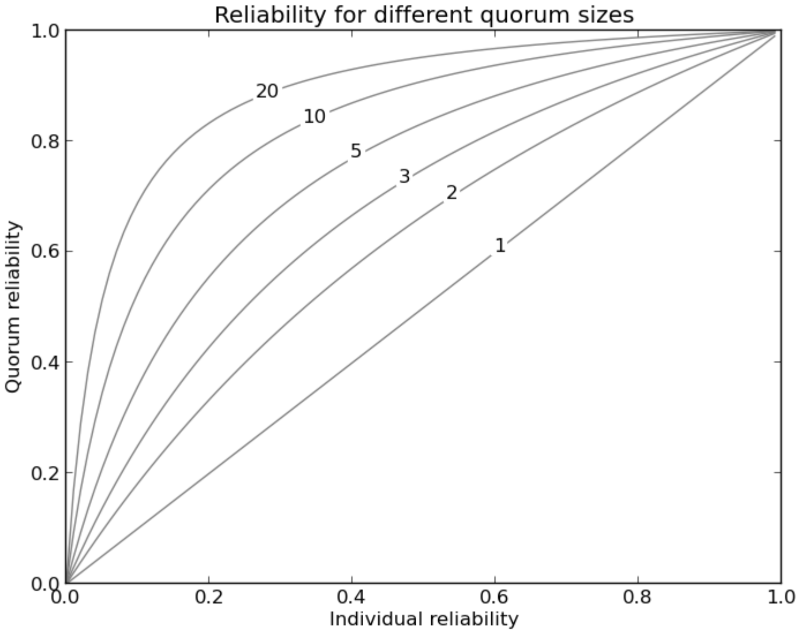
$$R_k = kR_1 / (kR_1 - R_1 + 1) \tag{3.16}$$

Quorum coding has some useful properties. First, as the size of the quorum grows, inter-quorum reliability improves. In the limit of an infinitely large quorum, inter-quorum reliability approaches 1—a perfect score.

Another elegant feature of quorum coding is that it does not require extra data to estimate the effective reliability of a quorum. If we can assess intercoder reliability among individuals, we can calculate intercoder reliability among quorums of any size. Figure 3.3 illustrates the improvement to intercoder reliability among quorums as the size of the quorum increases. The x-axis denotes reliability among individual coders, and y-axis gives the estimated reliability among quorums of size 1, 2, 3, 5, 10, and 20.

The fourth column of table 3.3 applies this reasoning to the items in our codebooks. It gives estimated reliability scores for quorums of size five. Most of the estimated group scores for our codebook items are in the range of .6 to .8, scores considered quite strong by the usual standards of content analysis. The lesson is that small groups of novices can effectively match expert-level reliability. Whether or not they are coding for exactly the same things is another question, which I will take up in section 3.4.5.

Figure 3.3: Quorum reliability curves



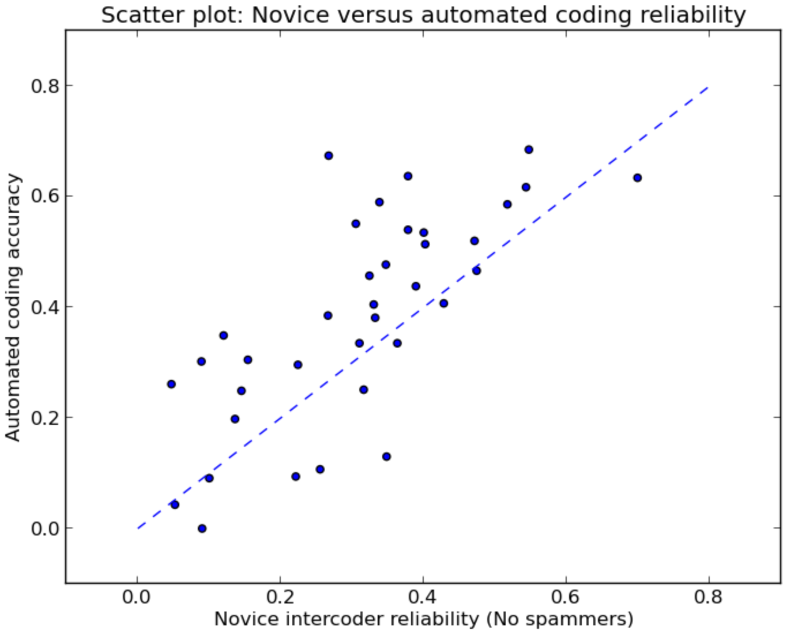
3.4.4 Automated coding

From the perspective of text classification, all of the work we've done so far has been to label documents and train classifiers. After all this work, the crucial question is, "How accurate are those classifiers?"

To assess accuracy, I took advantage of the benefits of 5-by-25 replication and quorum coding. I did this by training the classifiers on articles that were only coded once, and then assessing classifier reliability against quorum averages on documents that were coded multiple times. To be specific, I treated the text classifier as one "coder," and treated the quorum average of up to five human coders as a second "coder," then calculated intercoder reliability between the automated coder and the group coder in the usual way. This approach allows us to assess classifier reliability without penalizing the classifier as much for noise and mistakes in the novice coding[13].

[13] To be even more specific, I used a random sample of 90 percent of the available five-coded documents as a training set. I then took the remaining 10 percent of five-coded documents, appended them to 90 percent of the once-coded documents, and used the joined set as my testing set. This approach preserves the notion of comparison to a quorum while still allowing cross-validation.

Figure 3.4: Scatter plot: Novice intercoder reliability versus automated accuracy



The dashed diagonal line has slope of one and passes through the origin. For points above this reference line, automated coding was more reliable than novice coding.

This approach turned out to work reasonably well. On most of the codebook items, text classifiers outperformed the novice coders in terms of intercoder reliability. Where novice coding was in the .20's, the classifiers typically scored in the .3 to .5 range. For items with better novice scores, the reliability for the text classifiers often reached into the .4's, .5's, and .6's.

Figure 3.4 shows a scatter plot of novice and classifier intercoder reliability scores. This graph tells two main stories. First, novice and automated intercoder reliability scores are correlated across items. Second, automated scores are typically higher than the scores for individual novices. This evidence suggests that the statistical methods for text classification tend to balance out the idiosyncracies in different human judgements. They pick up on the signal in the training data, and filter out most of the noise.

These numbers also provide strong empirical support for the theoretical argument around quorum coding made in section 3.4.3. Our classifiers are trained on individual coding, and evaluated against quroum coding. If quorums were not more reliable than individuals, it would be almost impossible to obtain higher accuracy scores from classifier testing.

3.4.5 Comparing automated and expert coding

The coding system outlined in this chapter is a chain with two links among three kinds of coders: expert to novice, and novice to automated. Thus far, we have tested the second link and found that it holds firm: on virtually every codebook item with reasonably strong novice intercoder reliability, the classifier was able to meet or exceed that level of accuracy. However, it remains to be seen whether novice (or automated) coding correlates with expert coding. Although we have seen that quorums of novices exhibit reasonably high intercoder reliabilty, this only proves that novices have reasonably high internal reliablilty among themselves. It is still possible that the novice coding differs in important ways from expert coding. In this final results section, I carry out an indirect test of the expert-to-novice link in our chain of evidence.

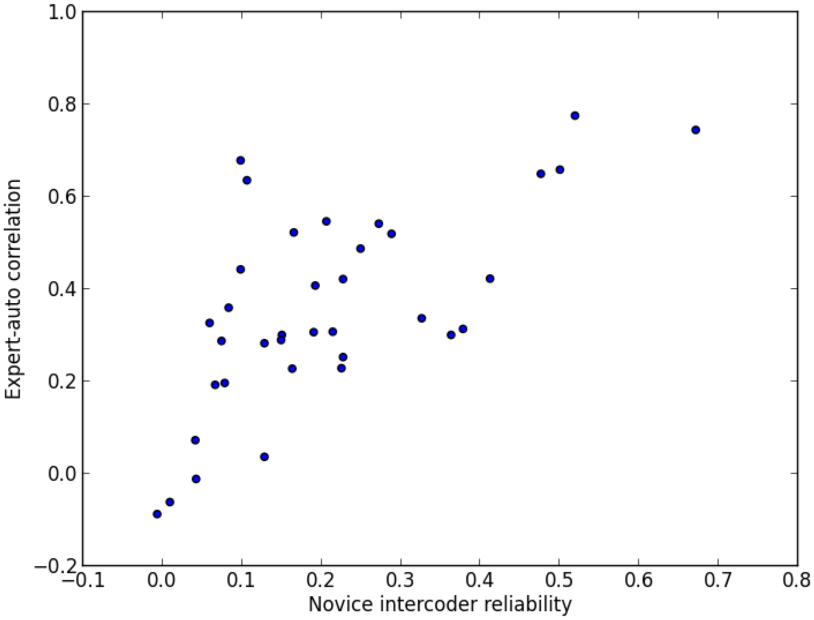
Unfortunately, our data do not allow for direct tests of correlation between novice and expert coders. Although each group coded comparable documents from similar samples, they did not actually code the same documents[14]. Consequently, we must rely on a less direct and more difficult test: examining correlations between expert and automated coding. Essentially, this approach tests both links in the chain at the same time. If expert and automated coding turn out to be correlated, we can infer that novice coding is also correlated with expert coding.

As shown in Figure 3.5, two main results emerge from this analysis. First, many of the correlations between expert and automated labels are respectably high[15]. These

[14] See Section 3.3.5 for details.

[15] Exactly how strong these correlations must be to be considered "strong enough" is a tricky

Figure 3.5: Scatter plot: Novice intercoder reliability and correlation between automated and expert coding, by codebook item



are the points with high values on the y-axis. For these strongly correlated codebook items, we can be confident automated coding reflects expert coding, and vice versa.

Second, the strength of the correlation between experts and automated coding is generally at least as strong as the novice intercoder reliability score. That is, there are no points far below the unit line. If novices were coding for different constructs than experts, we would expect to see points well below the unit line. The fact that most codebook items have stronger expert-to-automated correlations than novice intercoder reliability is moderately strong confirming evidence that the expert and novice coding are correlated as well.

question, related to the question of what levels of intercoder reliability are acceptable for measurement. I take up this question in section 3.5.1.

3.5 Discussion

Stepping back, we now have evidence to assess the effectiveness of this methodology and relate what we have learned to the practice of large-scale content analysis more broadly. Accordingly, this section proceeds as follows. The first section compares reliability across coding methods in order to draw inferences about the strengths and weaknesses of this methodology. Next, we review assumptions and threats to validity, focusing on the possibility for bias in our measures. The third section is an epistemological discussion of ground truth in content analysis—an argument for using novice coding to ensure transparency and replicability. Last, we discuss implications and avenues for future work in this area.

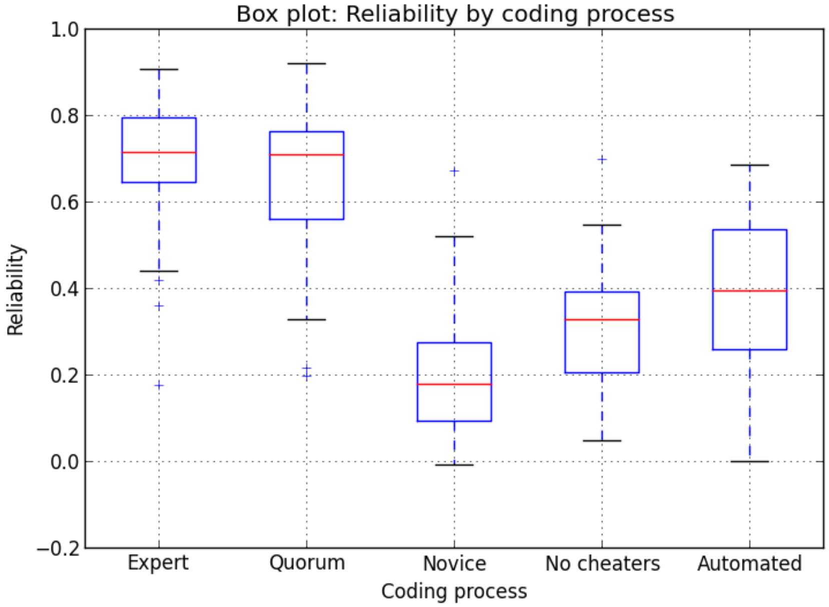
3.5.1 Comparison of coding methods and codebook items

This chapter has focused on three coding methods: expert coding, novice coding (and its extension, quorum coding), and automated coding. Setting other considerations aside for the moment, it is reasonable to ask which of these methods is the most reliable.

Broadly speaking, expert coding and five-fold quorum coding were the most reliable methods, with average intercoder reliability scores of .697 and .642, respectively. These methods are followed by automated coding, where intercoder reliability scores averaged .385 across all codebook items. Individual novice coding was the least reliable, with average intercoder reliability of .210 before excluding spammers, and .309 afterward. Figure 3.6 illustrates this pattern with box plots of item-level reliability scores, grouped by coding process.

As a a stylized fact, it is fair to say that among the 37 codebook items tested here, the intercoder reliability of quorums of five novices was comparable to experts. If we take a Krippendorff's alpha score of .5 as our cutoff level for acceptable reliability, 33 measures pass the test under expert coding, as opposed to 28 under quorum coding.

Figure 3.6: Boxplots of item-level reliability for experts, quorums, novices, and automated classifiers



Each set of box-and-whiskers illustrates the interquartile range for the 27 cookbook items used in further analysis. Outliers are shown using the "+" symbol. In general, quorums of five novices achieve comparable scores to experts, and automated coding outperforms individual novices even after spammers are been removed.

Table 3.4: Codebook item counts by coding method and reliability cutoff threshold

Threshold	**Expert**	**Quorum**	**Automated**
.3	36	35	25
.5	33	28	12
.7	21	20	0

On average, expert coding outperforms quorum coding by a small margin.

However, averaging across items oversimplifies the relationship somewhat. If we compare reliability on a per-item basis, we find quorum coding has higher intercoder reliability than expert coding on 20 out of our 37 codebook items[16]. This pattern is largely explained by a small handful of codebook items with very poor novice reliability (e.g. “non internet info, ” open-minded, reasonable person,). These items drag down the average reliability of quorum coding enough to put expert coding in the lead[17].

Table 3.4 shows the number of codebook items that are acceptable given different reliability thresholds. At a somewhat forgiving .5 cutoff level, 33 items passed the test for experts, and 28 for quorums. Even at a relatively stringent .7 cutoff level, 21 and 20 items (respectively) pass. From a research design perspective, the codebooks were designed with some redundancy in mind. It isn't important for *every* codebook item to be highly reliable, as long as we can accurately measure the underlying theoretical constructs. For both expert and former coding, these results are quite encouraging.

Results for automated coding are more mixed. Recall that our text classifiers were trained on once-coded novice data, and evaluated against five-coded quorum data. Measured against individual novice coders, the classifiers show up very well. The

[16] Careful inspection of table 3.3 reveals that quorums outperformed experts on the neutral voice, public import, and respect for others codebooks. On the other hand, experts did better on information sources, sources and evidence, and divisiveness. Within the appeals to fear codebook, performance was mixed. Despite careful searching, I can find no clear predictive patterns for when quorums outperform experts. The subject matter, question wording, and answer categories of codebook items all appear to be unrelated to expert-versus-quorum performance. In light of these non-findings, it seems that the most likely explanation lies in the mix of experts and novices who supplied data for each cookbook.

[17] Note that we could easily push quorum coding into the lead by replicating coding more times. The decision here is really one about acceptable cost: how much redundant content coding is feasible, given the constraints on our research budget?

average alpha score for automated coding was .385, compared to .309 for novice coders with spammers removed. Classifier reliability exceeded individual-level intercoder reliability for 26 out of 37 codebook items. This echoes results from Chapter I: text classifiers are capable of synthesizing “the wisdom of crowds” well enough to outperform individual humans. In Chapter I, we evaluated a single measure—political content—against individual novices. Now we have replicated this finding across several dozen measures, with evaluation against quorum coding so that we can assess the difference between classifiers and individual humans even more accurately.

At the same time, the coding tasks in this chapter are considerably more nuanced than the simple “political/not political” distinction made in Chapter I. As a result, the reliability of individual novices is considerably lower. Therefore, even though automated coding outperforms novices, it still often falls short of ideal levels of accuracy as measured on an absolute scale.

This begs the question, “When it comes to intercoder reliability (and classifier accuracy), how high is high enough?” Sadly, the literature does not provide a clear answer. Krippendorff offers .800 as a desirable standard (“Rely only on variables with reliability above alpha = .800.”), and .667 as a lower bar for “drawing tentative conclusions.” However, Krippendorff seems reluctant to set absolute standards (“Although every content analyst faces this question, there is no set answer.” “I recommend such levels with considerable hesitation.”). Moreover, his theoretical requirements seem to be much more stringent than what most researchers (and reviewers) use in practice. Much of the content analysis literature fails to report intercoder reliability at all. The literature on crowdsourced data collection is similarly ambiguous. Much of the research in this field reports accuracy, precision, or recall statistics, which are sufficient to compare methods within a single experiment, but provide no basis for objective comparison across experiments[18].

[18] When it is possible to calculate measures of reliability, such as Krippendorff’s alpha, the results often turn out to be quite low. For example, a widely cited paper by Shaw, Horton and Chen (2011)

If we take a relatively stringent .7 score as our cutoff line, then none of the automated coding measures are acceptable—a disappointing result. If we step back to a somewhat more forgiving .5, then 12 automated measures pass muster. If we relax our conditions further to .3, then we are left with 25 usable measures[19].

The decision over what makes an acceptable standard for intercoder reliability is complicated by the fact that the appropriate level of reliability depends on the analysis in which the measures will be used. As Krippendorf points out,

> some content analyses are robust in that the unreliabilities that enter the data-making process are barely noticeable in the results. In others, small differences may tip the scale in important decisions. ... To appreciate this sensitivity, analysts would have to know how disagreement in the data is transmitted through the analytical process to its outcome."

Ultimately, the necessary precision of a measure depends on the effect size one is hoping to estimate.

3.5.2 Measurement error and validity

This brings us to the issues of measurement error and downstream validity. Two questions are intertwined here. First, what kinds of measurement error are problematic for further analysis? Second, what kinds of error are likely to emerge from this methodology? These two questions are at the core of the emerging discipline of text-as-data analysis.

This section is organized around answers to the first question. I discuss three types of error that could distort downstream results: noise, uniform bias, and conditional

reports accuracy statistics ranging from 25.6% to 73.2% for five tasks. The authors do not report Krippendorff's alpha scores, but we can calculate the based on administrative data in the paper. Intercoder reliability scores range from .642 to -.048, and are *negatively* correlated with accuracy.

[19] Admittedly, .3 is a very low bar for reliability—more signal than noise. We will discuss implications for large-scale analysis in the next section.

bias. My treatment of these errors is drawn from the traditional econometric and statistical literatures on measurement error. Along the way, I will highlight examples within the context of content analysis.

By *noise*, I mean variables where errors in document codes have expected values of zero, and are strictly independent. In bivariate OLS and related models (e.g. Pearson's correlation, large-sample students t-statistic), noise attenuates estimates towards zero. This property of noisy variables makes statistical tests inherently conservative. In multivariate OLS, noisiness can have secondary effects on other parameter estimates, conditional on the correlation matrix. As a rough heuristic, we might say that items with low intercoder reliability are safe as the objects of direct study, but make for poor control variables.

In content analysis, noisy variables are ubiquitious. The reliance on subjective human coding and resulting focus on intercoder reliability make measurement error difficult to ignore. Among novice coders in particular, incentives to code as many documents as quickly as possible probably make noisy coding especially prevelant. For this study, the statistical characteristics of noisy variables come with a note of optimism. For very large-scale analysis, we can often aggregate and average across large corpora of documents to cancel out random errors. If this is the case, then even measures with fairly low reliability at the document level could yield accurate, valid estimates at the population level[20]. However, this approach depends on the assumption that the measures are simply noisy, and not systematically biased.

I define *uniform bias* as errors in document codes that have nonzero expected mean, but are otherwise independent of each other and other variables. Somewhat surprisingly, measurements of this type are unlikely to have much impact on most analysis. This is because most measures in content analysis are essentially ordinal, not cardinal: we care about the relative ordering of documents on various dimensions,

[20] In many ways, this is the core intuition behind Hopkins and King (2010).

but do not need to place them on an absolute scale. In fact, given the inherrent subjectivity of civility and newswriting, it is difficult to see how they *could* be placed on an absolute scale. As a result, most of the analysis we would wish to perform will be constrained to comparisons of means and correlations—operations that are unaffected by additive bias in our measures.

For example, suppose that blog posts in our sample average a score of four according to some measure of divisiveness. Suppose further that novices' codes for divisiveness are biased upwards, with an average of six—perhaps the novices believe that our study is looking to find that divisiveness is widespread, and obligingly incline their answers in that direction. Is this a problem? It is if we want to make claims such as, "The average blog post scored a four for divisiveness." But since the answer scale here has no external reference, this kind of statement is meaningless. Instead, we are more likely to make claims such as, "On average, Democrats write blog posts that are .3 standard deviations more divisive than Republicans," or "Divisiveness shows a slight correlation with blogger age (r=.14)." These kinds of comparative statements are unaffected by uniform bias. In other words, in most cases, uniform bias is only a problem for analysis we would not conduct on subjective measures anyway.

In its pure form, this argument only holds for unbounded continuous measures where bias affects the mean and not the variance of the measure. For measures with a small range (e.g. five-point ordinal scales), bias could influence not only the mean but the variance. However, even in this case, we have reason to be optimistic. If the variance of a variable is affected by bias on a constrained range, the effect will almost certainly be to decrease the total variance of the measure. The likely outcome will be for statistical analysis based on the variable to exhibit a conservative biased towards zero—not ideal certainly, but less problematic than bias in the other direction. A second line of defense comes from codebook items themselves: many items are distributed around the middle of their answer scales[21]. To the extent that this is true, the size of possible bias due to a constrained answer scale is limited.

Another surprising feature of uniform bias is that systematic bias by individual coders (e.g. as in Welinder et al., 2010) does not materially affect our results, as long as the coders do not all share a common bias, and each coder's contributed coding is small relative to the total set of labels supplied by the population. This is because assignment of documents to coders is strictly random, and therefore uncorrelated with any other variables of interest. Over many coders, any personal bias will average out. In effect, individual biases in a large population are actually just a source of noise.

Conditional bias, where document errors have nonzero mean conditional on some other variable, is much more problematic. In this case, OLS regression and any of its derivatives can be subject to omitted variable bias. For example, if we recruited an overwhelmingly Republican group of novice coders, we might discover that they naturally code blog posts by Republicans as more civil and blog posts by Democrats as less civil. Any subsequent analysis of the relative divisiveness of Republican and Democratic bloggers would be biased in favor of Republicans. Even worse, any analysis of divisiveness alongside other variables correlated with partisanship would also be biased.

Unfortunately, no simple test can exclude the possibility of conditional bias. Indeed, if we attempt to absolutely rule out bias with respect to any conditioning variable, we find ourselves in the impossible position of trying to disprove a negative. Even so, we can take certain steps to reduce the possibility of conditional bias in novice coding. Here I suggest three simple precautions to reduce the risk of conditional bias.

Don't disclose the purpose of the study in instructions. This is standard practice in most survey research and psychological experiments (King, Keohane and

[21] This happens for the same reason that many survey items tend to be centered in the middle of their scales: researchers deliberately write them that way, because this measurement property simplifies downstream analysis.

Verba, 1994). It seems reasonable to employ the same precaution in content analysis.

Draw from a "normal" pool of coders. Many of the same arguments used to defend the notion that college sophomores are a good sample population for psychological experiments can also be applied to content analysis on MTurk. Although the population isn't strictly representative of any larger population, there is little reason to believe that MTurkers are cognitively different from "normal" adults (Buhrmester, Kwang and Gosling, 2011, Berinsky, Huber and Lenz, 2011).

Compare novice coding to experts. In the next section, I will argue that novice coding should be preferred over expert coding as the source of "ground truth" in content analysis. Still, comparing novices to experts can provide a useful validation step for measurement. When data collected by experts and novices are strongly correlated, it becomes harder to argue that conditional bias is at work[22].

Let us conclude this section with a last comment about conditional bias and subjectivity. First, the notion of conditional bias becomes complicated in the context of subjective coding. Take the example of Republican coders rating Democratic bloggers as less civil. From the subjective perspective of the coders, those judgements are correct. Other coders might disagree, of course, but that is the nature of subjective judgement. It seems strange to call it "bias" when the Republican coders are simply giving us what we asked for.

As I will argue in the next section, a better approach is to treat the Republican coders as part of a quorum of novice coders. From this perspective, the conditional "biases" of each individual coder are actually decision-making heuristics that inform the collective judgement of the quorum. From an epistemological standpoint, I find this rationale coherent and satisfying. From a methodological standpoint, I suspect

[22] This was an notion behind my indirect tests in section 3.4.5. Future researchers would do better to learn from my mistake and collect data to allow *direct* comparisons.

that embracing diverse heuristics in this way will make subjective content coding a more robust and accurate approach to measurement.

3.5.3 Ground truth in content analysis

In the course of this chapter, we have introduced three different ways of measuring variables from content. Among expert, novice, and automated coding, each has different virtues in terms of reliability, replicability, and scale. But of course, they are not identical. We turn now to an important epistemological question: which measure is the right measure?

In this section, I will make the case that in subjective coding tasks, quorum coding by novices deserves to be treated as the "true" measure. To make this case, we will first define subjective coding tasks. Next, we will consider the relative merits of expert and novice coding. Finally, we will turn to the relationship between automated, novice, and quorum coding.

Let us begin by drawing a distinction between "objective prediction tasks" where the truth can be verified (possibly after some delay) against an objective oracle, and "subjective judgement tasks" where the only source of truth is human perception. For example, guessing the weight of a cow at a county fair, or the atomic weight of cadmium, or the value of the NASDAQ next week are all prediction tasks. Responses to the statement, "The author shows respect for opposing viewpoints" are much more subjective, since "respect" is an abstraction that is only meaningful in the eye of the beholder (or a crowd of beholders).

For prediction tasks, the right combination of coders is an empirical question. By measuring past performance against the truth as revealed by the oracle, we can estimate variance and bias by coders and coder types. With estimates for those quantities in hand, we can pick the optimal balance of experts and novices (Lamberson and Page, 2012). Even if we can't measure the exact variance and bias for each predictor, we can still make some educated guesses about the composition of the optimal crowd.

For subjective tasks, we don't have the ability to measure past performance, because the measures are inherently subjective – there is no oracle to which we can directly appeal to assess accuracy. In this context, the word "prediction" doesn't really make sense. "Judgement" or "perception" is more apropos, because the "correct" rating is a subjective, personal decision. Even the definition of "expert" and "novice" changes when we go from objective tasks to subjective tasks. In objective prediction tasks, an expert is usually one who has demonstrated high accuracy on past prediction tasks, or has special training or experience that we expect will make them more accurate (Ashton, 1986). In the realm of subjective tasks, an expert is simply one who was involved in the design of the codebooks—a coder who happens to know the intent of the research design and interact with the other research designers.

Given the choice between expert coding and coding by novices, it's tempting to say that the expert coding is the "real" coding. After all, the experts are the ones who created the codebooks[23]. This approach becomes even more tempting when (as is usually the case) the expert coders have higher intercoder reliability.

However, if we believe in the scientific virtue of replicability, then novice coding is actually superior. The scientific method is fundamentally democratic, based on the premise of transparency and replication—any observer should be able to replicate an experiment and obtain the same result. Scientific arguments are supposed to be grounded in measurement and evidence, not authority and expertise. In the case of content analysis, this means that we should certainly prefer novice coding over expert coding, because only novice coding is a transparent, replicable process.

Results in this chapter demonstrate that codebook items with high intercoder

[23] Actually, this awareness of the purpose of the study gives us another reason to avoid expert coders. With novices, we worry that they *might* guess the intent of the study, and thus bias results. With experts we are certain that they *already* know the purpose of the study. If our goal is to avoid corrupting our research through biased coding, the choice of novices over experts seems clear.

reliability among experts can still generate poor reliability among novices. Such items stand as cautionary examples for human replication. They warn us that just because a tight-knit group of colleagues are able to agree on document labels, we should not yet be persuaded that others outside the clique can do the same. Instead, each item creates its own implicit hypothesis test: "A small team of experts was able to code this item with high reliability. Can novice coders do the same?" Very few content analysis studies validate this hypothesis.

To summarize: in this study, we have good circumstantial evidence that the items in our codebooks were coded in similar fashion by experts and novices alike. However, supposing that differences between experts and novices had emerged, it would have been better to treat novice coding as the "real" data source. Given a choice among measures, the scientific method dictates that we should use the measurement that is replicable.

What about automated coding? Automated coding is supremely replicable, and has the added virtue of scalability. In most cases, automated coding is actually more accurate than individual novice coding. On the other hand, automated coding isn't a single process; it's an infinite family of possible processes. To see this, consider the many decisions made in the course of training a text classifiers: which algorithm to use, which features to include, what parameters to use when training the classifier, etc. Different choices at any stage of the training process will lead to different classifier features and weights, and therefore different classifier values. Unless we are willing to accept all of these potential classifiers as equally valid, we must have some additional criterion for judging among them. In the case of content analysis for subjective constructs like newswriting and civility, the obvious candidate for that criterion is replication of human judgement.

However, there's some epistemological subtlety to making a claim like this. I have claimed that (1) human coding is ground truth for text classifiers, and (2) text classifiers are often more accurate than human coders. This sounds like a contradiction, but it can be solved by differentiating between individual and quorum coding. Specifically, I define ground truth for X as the *average* label assigned by an infinitely large quorum of novice coders.

$$X = \lim_{k \to \infty} X^{kH} \tag{3.17}$$

Other researchers have converged on similar definitions. Conway (2013) also uses quorum coding on MTurk to measure political constructs in text. Conway makes heavy use of group averages for inference, and although he generally treats expert coding as ground truth, he is also critical of known biases and flaws in expert coding. In his dissertation manuscript, Zhou (2013) advanced a similar notion of "distribution as ground truth." In this approach, which is grounded primarily in the discipline of machine learning, "human coders are not merely an instrument; rather, the coders, who represent the population if randomly sampled, collectively define the ground truth of items as distributions by aggregating subjective opinions."

This definition has many useful properties. Because it is based on novice coding, the measurement process is replicable and, to some extent, scalable. Although we can never measure X perfectly, we can use repeated coding to approximate it to any level of statistical precision we choose. Moreover, we can use small samples of novice coders and the logic of quorum coding to forecast the necessary level of repeated labeling. When machine learning algorithms are sufficiently accurate, we can use automated coding as a highly scalable proxy for quorums of novices.

Now, this might feel dangerously close to abandoning scientific measurement altogether and just going into politics (”The truth is whatever the majority claims it is.”) Actually, it's a call for rigor and replication in content analysis. My claim is that in the domain of subjective tasks, we must replicate coding with novices in order to demonstrate validity (”The truthiness of a measure depends on our ability to replicate it with high quorum reliability.") Given the choice between expert and novice coding, we should prefer novice coding, because only novice coding satisfies the essential scientific criteria of replication[24].

3.5.4 Lessons for large-scale content analysis

As mentioned earlier, I am convinced that large-scale text-as-data research will be very fruitful in coming years. As more and more texts become available in digital form, the potential for research using the methods of content analysis will increase dramatically. At the same time, I expect that generally applicable principles for developing reliable codebooks and techniques for scaling coding to large corpora will be studied further and pass into widespread use. The resulting combination of replicability and scalability, plus the flexibility that comes from human judgment, will become very powerful. Similar to the explosion in survey methodology 70 years ago, large-scale content analysis seems poised for growth as a mainstream scientific methodology.

However, as we have seen in this chapter, large-scale content analysis has many moving parts, and all of them have to work well in order to arrive at valid conclusions. Accordingly, I close with thoughts on three areas of focus for productive research within this emerging field.

1. Design and build better systems for recruiting, training, and incentivizing coders. Within certain niches of social science, small research teams have been conducting content analysis for many years (Neuendorf, 2002). This cottage industry has made important contributions in several fields (e.g. Baumgartner et al., 2003). However, with few exceptions, these teams have lacked the tools for large-scale content

[24] The way I've phrased the conclusion ("novice coding trumps expert coding") is a bit of a false dichotomy. From the perspective of minimizing error, it seems reasonable to combine expert and novice coding together, as in Lamberson and Page (2012). However, bear in mind that novice coding isn't just a substitute for expert coding; it's an essential step for validating expert coding. We should be very skeptical of subjective expert coding that has not been replicated by novices.

labeling, and have sometimes neglected replicability and external validity in the process of data collection (Mikhaylov, Laver and Benoit, 2008).

The advent of online crowdsourcing platforms opens the door to remedy these shortcomings. However, platforms like Mechanical Turk are still in their early stages and suffer from poor design decisions, usability problems, and incentive mismatches (Ipeirotis, Provost and Wang, 2010). Within computer science, a sizeable literature has sprung up seeking to alleviate these problems (e.g. Karger, Oh and Shah, 2011; Sheng, Provost and Ipeirotis, 2008). Although the quality of research in this area is uneven (Adar, 2011) and we are still far from full solutions (Kittur et al., 2013), I am optimistic about the general direction that this literature is going.

With that said, much of this research seems surprisingly unconcerned with the human aspect of crowdsourcing. Although systems for distributing and aggregating tasks abound, well-designed studies of incentives, communication, and learning are surprisingly hard to find (Kittur et al., 2013). Given the essential role that human perceptions and responses play in crowdsourcing, it seems reasonable to suppose that the field will need to come to grips with how human beings adapt within such systems.

Within the social sciences, there is growing enthusiasm for MTurk as a platform for surveys (Buhrmester, Kwang and Gosling, 2011; Berinsky, Huber and Lenz, 2011) and experiments. Given the prevalence of surveys and experiments within social science, this makes sense. However, the primary intended use-case for MTurk is repeated tasks by the same workers—a perfect match for content analysis. I have been surprised at the slow rate of adoption among content analysts, but expect that to change before too long.

When that happens, we should expect greater synthesis between theories and methods from social and computer science. Going out on a limb, I would venture that the microeconomic literature on personnel economics will come to play an important role in this literature, for two reasons. First, labor economists already have a great deal to say about rational approaches to recruiting, incentivizing, and retaining a labor force (Lazear, 1995). The theoretical models are mature, and have been tested—to a point—in the setting of human resources in traditional labor markets . Second, crowdsourcing platforms provide an ideal laboratory to test such theories (Horton, Rand and Zeckhauser, 2011), because it is one of the few labor markets that is sufficiently malleable to conduct such experiments. To date, few if any crowdsourcing studies deal with incentive structure with a level of sophistication that labor economists would recognize (e.g Mason and Watts, 2010, Shaw, Horton and Chen, 2011), but there is no reason this state of affairs need be permanent[25]

2. Develop best practices for codebook development. After close to a century of public opinion surveys, it is easy to forget that the whole practice of survey research had to be invented, from scratch. Early survey researchers painstakingly tested different approaches to their craft, developed best practices that became standard training for future generations of researchers: standard answer scales (e.g. Likert, semantic differential), heuristics for question wording ("Avoid double-barreled questions."), sampling strategies for door-to-door and telephone surveys, psychometric procedures for creating valid scales from multiple questions, whole batteries of questions with standard wordings and known measurement properties (Fowler, 2009; Weisberg, 2009). This degree of attention to measurement and replication was a huge boon to the field—a scientific necessity, really.

By all appearances, we are at the same point with content analysis, transitioning from a cottage industry to a mainstream methodology. As we do so, I strongly suspect that we will need to develop a parallel body of heuristics and best practices, supported by appropriate statistical tests. These best practices may share common elements with survey research, but they will certainly need to be adapted to the unique needs

[25] The bonus structure used in my MTurk asignments was originally intended as an experiment with quality-based variable pay. In the interests of time, I set it aside. However, this is a project to which I certainly intend to return in future work.

of content analysis. Without this cumulative knowledge, I doubt that content analysis can become an effective tool for scientific measurement[26].

Let me propose two examples from the codebooks in this chapter. First, my results suggest that codebooks for content analysis should reference the text, not the author. Among my codebook items, many of the ones that performed poorly were those that mentioned the author (e.g. "The author is...," "The author seems to...") or worse, the author's intentions (e.g. "The author tries to...," "The author sounds like..."). Apparently, asking coders to speculate about authors and their intentions confuses the coding process.

Second, my results suggest that examples in direct quotations are very helpful for coders. For example, "Much of this text is spent criticizing others motives (e.g. 'he's out to get us,' 'she can't be trusted,' 'greedy,' 'dishonest,' etc.)" and "The author talks about threats to personal economic interests (e.g. jobs, income, tax rates, etc.)." I only included examples of this type for a handful of items, but these were among the best-performing items in all the codebooks—some of the places where novice reliability compared most favorably to expert coding. Rather than defining words with more words, using examples drawn directly from text seems to be an effective strategy for reducing ambiguity and improving intercoder reliability.

I put these two patterns forward as hypotheses. They seems reasonable, especially given the post-hoc supporting evidence from this study, but further research is needed before we draw strong conclusions.

Fortunately, both of these hypotheses could be tested directly by experiment. Research of this kind would be a productive addition to the literature on content analysis and crowdsourced data collection. It would build up a cumulative set of best practices for the discipline, and greatly reduce the trial-and-error burden on researchers and practitioners. Over time, I expect that the field will follow survey

[26] I note in passing, that such attention to measurement is largely missing from the field of sentiment analysis.

research by developing standard instruments for measuring important quantities, and standard corpora for validating new constructs.

3. Develop statistical methods that fully support the text-as-data research pipeline. Within machine learning, there has been considerable interest in algorithms that are robust to noisy labels (e.g. Prelec, 2004 Welinder et al., 2010; Zhou et al., 2012; Liu, Peng and Ihler, 2012). This is a good start. However, within the KDD literature, researchers have largely taken the challenge to be "replicate human coding as accurately as possible." However, for content analysis, the goal usually requires additional steps of statistical analysis. As we discussed in section 3.5.2, replication of human coding is typically not enough to guarantee valid results in downstream analysis. As Krippendorff points out in his notes on sensitivity analysis, we must appreciate how noise and uncertainty propagate from coding through analysis.

Once we understand these facts, we can design our algorithms and estimators with the specific goal of reducing noise and eliminating bias. In a sense, statistical techniques accomplishing this would be comparable to survey weights: re-usable methodology allowing us to efficiently communicate results from early stages of the data pipeline to the final analysis. To date, I have only encountered one example of a method that takes this step in text-as-data methodology: Hopkins and King (2010). Unfortunately, the method is limited to calculating proportions within a population.

The next step would be extending Hopkins and King's insight to methods such as OLS regression. For example, we could derive a least-squares estimator that would accept document-level codes and item-level reliability scores as input, and adjust estimates for beta values and standard errors accordingly[27]. Deriving, implementing, and applying such an estimator would be a non-trivial exercise, but well within reach

[27] In addition to statistical methodology, this approach rasises some potentially thorny epistemological issues: given a noisy variable, should we rely on the variable as measured, or impute results to an idealized, error-free version of the variable? An answer to this question will be necessary to interpret the result of such an estimation procedure.

of statistical theory and computational tools[28].

Another promising approach is to propagate information from text classifiers back *upstream* to inform the data collection process. Broadly speaking, such approaches would fall under the scope of active learning (Cohn, Ghahramani and Jordan, 1996). Early research by Sheng, Provost and Ipeirotis (2008), Karger, Oh and Shah (2011), and others shows how this might be accomplished in crowdsourced data platforms. I expect that future work will make such approaches even more efficient and cost-effective.

3.6 Conclusion

This chapter has been a long exploration of methodology, technology, and epistemology for large-scale content analysis. In contrast to most work in related fields, I have sought to implement an end-to-end data pipeline with multiple steps to integrate expert, novice, quorum, and automated coding. Doing so demonstrates that a combination of methods from traditional content analysis and natural language processing can enable valid analysis of very large-scale content analysis, even in the case of highly subjective constructs like civility and newswriting. It also demonstrates the complexity and difficulty of getting all the pieces to work in tandem.

This chapter makes contributions in two main areas. One contribution comes from the epistemological argument in Sections 3.3.2, 3.4.3, and 3.5.3. I have made the case that—for subjective coding tasks, at least—novice coding isn't just a substitute for expert coding; it's an essential step for validating expert coding. Despite the new availability of technologies like MTurk, very few studies in content analysis cross the critical bridge to replicate analysis with novice coders.

The second set of contributions is methodological. This chapter has introduced

[28] If we are willing to make a few simplifying assumptions (e.g. errors in coding are normally distributed), then a modified sandwich estimator for heteroskedastic consistency would probably suffice.

several practical methods for improving reliability and scaling up online content analysis. In particular, the combination of 5-by-25 replication, methods for filtering spammers on MTurk, proofs for estimating quorum reliability from individual reliability, heuristics for codebook wording, and training automated classifiers on novice coding. The individual pieces may be foreshadowed in existing research, but to the best of my knowledge, applying them together in this way is unique. I hope that these techniques for scale and reliability will find application in the broader fields of text analysis and crowdsourced data collection.

APPENDICES

APPENDIX A

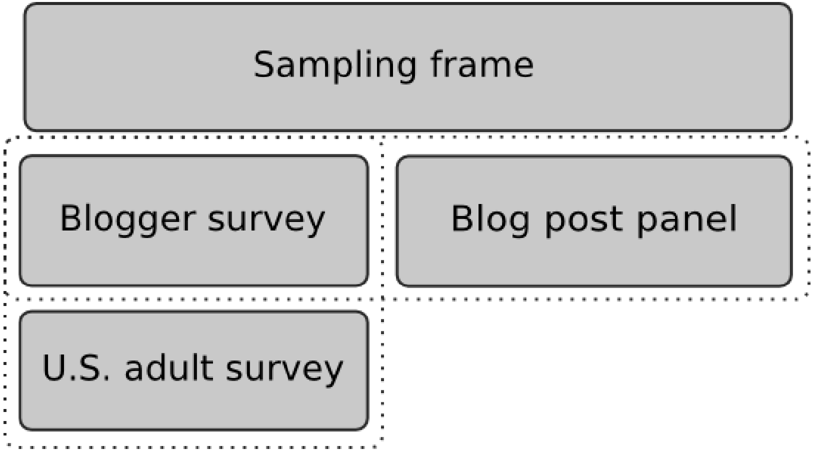
Data and Methods

A.1 What makes this research design difficult?

In the preface, I introduced a set of key questions about political bloggers and, more broadly, political activists. In order to answer these questions, we must sastify several rather stringent data requirements. First, in order to make claims about the blogosphere at large, our data must be drawn according to a representative sampling strategy. Second, our data set must include information about bloggers' attitudes, demographics, and social influences. Third, our data must also include observed content from online activism itself, in as much detail as possible. Fourth, our data must be collected in a way that allows comparisons to the general U.S. population, and other U.S. activists. Finally, our data must allow comparisons within the blogging population itself. Only a body of evidence with all of these characteristics can fully answer our intended research questions.

This appendix describes a data-collection process to fulfill those requirements, following three major steps: constructing a sampling frame, surveying political bloggers, and collecting corresponding content for their blogs. All three of these data elements,

Figure A.1: Research design schema



linked together, are necessary for the research design to be complete. The structure of this chapter reflects its three-part organization.

This appendix procedes as follows. The next section describes how to construct a sampling universe spanning the whole political blogosphere. Section A.3 details procedures for administering an opinion survey to a representative population of bloggers[1]. Section A.4 outlines a process for downloading and parsing a large panel of posts from the same blogs sampled for the survey. Finally, section A.5 concludes with a summary of the advantages of this research design, and a brief celebration of the potential of interdisciplinary research.

A note on technical material: this appendix touches lightly on technical issues such as algorithm design and sample weighting. My goal is to present enough detail for readers understand the chain of evidence so they can assess the validity of the arguments, without bogging down the argument with technical specifications. Readers interested in replicating these methods can find full details in the other appendices.

[1] This section is an abbreviated version of Chapter I.

A.2 Sampling strategy

A.2.1 Scope and definitions

The population of interest for this study is authors of *active, English-language blogs about politics* in the United States. I define content "about U.S. politics" more precisely as content *discussing what government in the United States is doing or should be doing.* At a practical level, this definition is close to common intuition about what constitutes political blogging. It includes content about elections, public opinion, public policy, public officials, and so on. State and local politics are included, as is U.S. foreign policy. So is content that mixes politics with other topics, such as religion, humor, or pop culture. Content mainly about politics outside the United States is excluded. At a theoretical level, this definition is closely related to the classical concepts of exercise and control of the authority of the state. It sits at the intersection of Hobbes, Key (1961), and Habermas (1991), and interfaces nicely with most existing work on representation, persuasion, opinion formation, discourse in the public sphere, and so on[2].

Following past research (e.g. Drezner and Farrell, 2008; McKenna and Pole, 2008), I define blogs in terms of format: a blog is *a web site on which the main content is a series of posts displayed in reverse-chronological order.* An *active* blog is one with at least one post in the last six months.

A.2.2 Sampling universe

In the past, no complete index of political websites existed. Lacking a sampling frame, previous studies of political blogs have been based on convenience samples

[2] Of course, other definitions of political content are also possible. We might include any content that is "about power relationships," or "evokes political topics in readers' minds." These are perfectly valid theoretical definitions, but in practice I have found that such definitions are hard to measure accurately. The first definition is subject to widely different interpretations, and therefore seems likely to be unreliable. The second definition is even more problematic, because it locates the definition in reactions to content, rather than content itself.

of one kind or another (e.g. McKenna and Pole, 2008; Davis, 2009). To solve this problem and create a representative sampling frame, I use a novel recombination of two technologies from computer science—automated text classifiers and web crawlers[3].

Automated text classifiers were originally developed by linguists and computer scientists working in the fields of natural language processing, machine learning, and information retrieval. The basic problem is to sort documents into predefined categories based on their contents. Categories can be defined in any way researchers choose, giving text classifiers a remarkably degree of flexibility. (See Manning et al., 2008 for a good introduction to text classification.)

Standard procedure in text classification is to hand-code a "training set" and "testing set" of a few hundred documents, calibrate the classifier on the training set, and then assess the level of intercoder agreement between human and computer coding on the testing set. I applied this procedure to train a classifier to recognize political content. Trained using text from 1,000 hand-coded sites, and evaluated against 200 more, the computer achieved higher intercoder accuracy and reliability than the human coders themselves. Human-computer agreement was 81.0 percent, slightly outperforming the 80.9 percent human-human agreement (Krippendorf's alpha = .733). This is impressive, but not too surprising, since a well-trained text classifier can learn to split the difference between human coding styles.

Web crawlers, also called *web spiders*, are computer programs designed to follow hyperlinks on the World Wide Web, scan web pages, and store interesting content. They are commonly used by search engines, such as Google or Yahoo, to catalog web sites.

To solve the problem of constructing a sampling universe for political blogs, I

[3] Studies on topical web crawling (a.k.a focused or directed web crawling) often combine classifiers and web crawlers as well (Menczer, Pant and Srinivasan, 2004; Chakrabarti, Van den Berg and Dom, 1999; McNamee et al., 2002). In general, these efforts have favored high precision over high recall, making them inappropriate methods for constructing sampling universes. See Noren's 2011-2012 food blog study for a research design more similar to mine (Noren, 2012).

have written software that combines text classifiers and web spiders to conduct an "automated snowball census" of the political web. Guided by the text classifier described above[4], this program exhaustively explores the World Wide Web in search of political content, using the following algorithm.

1. Start from a seed batch of political sites.

2. Download and classify each site in the batch.

3. For English-language political sites, harvest outbound hyperlinks and add unvisited links to the next batch.

4. Repeat from step 2 until no new links are found.

Any political website with at least one in-link from another political site in the snowball will be included in the final census. Since political sites are most likely to link to other political sites, this method rapidly and effectively charts the boundaries of the political web. Of course, sites without any in-links will not be found, but since such sites would not be found by search engines' crawlers either, they are effectively invisible and therefore outside the public sphere.

To construct the sampling frame for this project, I ran the snowball program in August 2010. It executed in 20 hours, crawling 1.8 million sites, of which 789,818 were political sites and some 42 percent were blogs. Based on the error rate of the classifier, these sites comprise an estimated 84.2 percent of the English political web. This sampling frame is far superior to anything existing previously, making it an excellent starting point for a representative survey of political bloggers.

Additional technical details for this census can be found in Chapter I.

[4] Note: I use an additional classifier to identify English text. Following standard practice, this classifier takes character trigrams as input and classifies with extremely high accuracy. In a testing set of 200 documents, all were classified correctly as "English" or "other."

A.2.3 Filtering and Stratification

The snowball software gives us a census of all political web sites. However, our sampling frame for this study is a specific subset of that universe: *active, English-language blogs* about *U.S.* politics. Filtering for these additional criteria takes some additional work.

Some of this work can be automated. As it happens, most blogs can be readily identified by the HTML tags used to format the site. Consequently, a simple classifier making use of those tags can identify blogs versus other sites with remarkable accuracy. Applying this classifier allows us to efficiently filter our list of "political sites" to down to "political blogs."

Next, we wish to draw a stratified sample, based on the popularity of political blogs. One useful side-product on the snowball census is a list of all the hyperlinks connecting political sites. Since the number of links directed to a site is a reasonably informative proxy for site traffic (Hindman, 2010), we can use this value to stratify political blogs by popularity. In network theory, this value is called the *in-degree* for a given site.

Based on site in-degree, I separated the sample into three rough strata: the top 5,000 political blogs, the next 45,000, and the remaining 422,000[5]. These values effectively capture very large differences in popularity. By oversampling on popular sites, we are able to increase variation on a set of key variables (popularity, traffic, level of participation, and level of professionalization) without ignoring the "long tail" of less active and less linked-to sites.

Following this stratification scheme, I drew samples from each of the samples. I drew approximately 5,000 sites from all three strata, with an extra 3,000 from

[5] These values are approximate. To save bandwidth and computation, I estimated in-degree cutoffs for each stratum (530 in-links for the first stratum; 213 for the second) before selecting samples to crawl. The net effect is that the exact counts for each strata were a little imprecise, but the strata still provide an accurate ranking of site popularity. For example, the top stratum turned out to contain 4,722 blogs instead of 5,000, but these are still the top-linked blogs in the census.

Table A.1: Sampling, eligibility, contacting, and response rates by stratum

	Strata 1	Strata 2	Strata 3	Total
Universe	$\sim 5,000$	$\sim 45,000$	$\sim 422,000$	$\sim 472,000$
Sampled	4,722	8,186	5,523	18,431
Eligible	2,248	3,389	1,216	6,853
Contacted	1,142	1,476	449	3,067
Responded	303	362	109	774
Percent eligible	47.6	41.4	22.0	37.2
Contact rate	50.8	43.6	36.9	44.8
Cooperation rate	26.5	24.5	24.3	25.2
Response rate	13.5	10.7	9.0	11.3

the second stratum, on the hunch that bloggers is this category would be roughly representative of the full population, but more likely to respond to my invitation. The exact counts for sampled sites by strata were 4,722; 8,186; and 5,523 respectively.

This "pre-screening pool" of 18,431 likely blogs was then handed off to a team of research assistants for final screening and collecting of contact information[6]. Each site was checked to verify that it met my criteria to be classified as a blog, had at least one post in the previous six months, and was focused on U.S. politics rather than politics in the U.K., Canada, Australia, India, or elsewhere. Table A.1 show results from this final wave of screening, as well as contacting and response rates. We will discuss these momentarily.

A.3 Survey

A.3.1 Survey administration

After sampling and filtering political blogs, the last major difficulty was gathering contact information. Many bloggers list contact information (especially email addresses) on their blogs, but this information is not listed in any standardized format:

[6] Some of this work was done using Amazon's mechanical turk, a popular crowd-sourcing service. See the recent paper by Berinsky, Huber and Lenz (2012) for discussion of applications of mturk in research settings.

even if an email address is there, it can be hard to find. During the filtering phase, the research team searched each blog thoroughly for an email address to contact the blogger. Email was my primary means for reaching out to bloggers.

When contacting bloggers, I followed best practices in survey administration as closely as possible. I sent a series of four emails inviting and reminding bloggers to complete the survey[7]. These short messages explained the purpose of the study in general terms (e.g. "This study looks at how the blogging community thinks about and responds to important issues and current events."), provided a link to the survey, and suggested a variety of reasons to participate: the survey is a chance to share ideas and experiences, we need everyone in the sample to participate to make sure results are representative, the survey is only 20 minutes long, etc. To the extent possible, I tried to personalize messages to make it clear that my invitations were not auto-generated spam. I also tried to invoke descriptive norms (Reno, Cialdini and Kallgren, 1993) and norms of reciprocity from social exchange theory (Dillman, 2007) by expressing appreciation, supporting group values, making the questionnaire interesting, etc. Appendix B includes the text of all contacting messages.

Because one of our goals was to enable comparisons between political bloggers and the population at large, I carefully replicated many aspects of the Evaluations of Government and Society Survey (EGSS). This nationally representative survey was conducted by the American National Election Study around the 2010 midterm elections. To reduce the threat of timing and mode effects in cross-sample comparisons, I conducted my survey at roughly the same time, in the same format (online, self-administered), and with many of exactly the same questions as the EGSS.

All told, I was able to gather contact information for 44.7% of sampled bloggers, and achieve a cooperation rate of 25.3% , with an effective response rate of 11.3% .

[7] To be more precise, the survey was administered in three waves with small differences in timing, incentives, non-email contacting, and contact messages. These differences had little or no measurable impact on survey responses, and are documented in Appendix B

These response rates are quite strong for unsolicited online surveys. It seems likely that much of the non-response is being driven by inactive email accounts.

In the end, this contacting strategy was successful, yielding a final sample of 774 respondents, drawn with known probability from the general population of active, English-language U.S. political bloggers.

A.3.2 Survey instrument

The survey itself was conducted online, deployed using the Qualtrics survey software package. In a series of pretests and cognitive interviews with a convenience sample of political bloggers, I revised and refined the survey instrument until it took most respondents about 20 minutes to complete.

Survey items were chosen with two purposes in mind: to explore the motivations of political bloggers, and to enable comparisons between bloggers and the population of U.S. adults at large. For this reason, I borrowed heavily from important "surveys of record" when writing the survey. Four out of five of my survey items are taken with little or no modification from the American National Election Study (ANES); Verba, et. al.'s classic study of political participation; and technorati.com's annual State of the Blogosphere.

The final questionnaire includes 337 items covering a wide array of topics: blogging habits, reasons for blogging, news and media consumption, attitudes towards groups and issues, political participation beyond the blogosphere, and demographics. Care has been taken to ensure that key variables from important empirical theories are all included: canonical theories of participation (Rosenstone and Hansen, 1993) (Verba, Schlozman and Brady, 1995), retrospective voting (Fiorina, 1981) and macropartisanship (MacKuen, Erikson and Stimson, 1989), Zaller's (1992) RAS model of opinion formation, network theories of social influence (Huckfeldt and Sprague, 1995) (Mutz, 2006), selective exposure (Sears and Freedman, 1967) (Prior, 2007). Measuring the same variables in the same ways as previous studies should enable future researchers to replicate important results from the field, an essential step in synthesizing these theories in the context of the political blogosphere.

A.4 Blog post panel

One of the most exciting opportunities in this project is our ability to observe the content of blogs alongside survey responses from their authors. I do so by downloading the full content of blogs in the sample, in panel format. All told, the full panel includes nearly 10 million blog posts, including full text, hyperlinks, and timestamps. This panel is an incredibly rich source of insight into behavior in the blogosphere—one of the largest and most detailed data sets on political speech and activism ever assembled.

A.4.1 Blog post parsers

Since their introduction in the late 1990's, blog posts have proven to be a remarkably stable genre. Virtually all blog posts include most, if not all, of six fields: body content, date, title, author, tags (or labels), and number of comments[8]. For downloading content, my goal was to capture these fields for as many of the sampled blogs as possible[9].

These fields can be extracted from any given blog by reverse-engineering the format of the blog and writing a small piece of software, called a *screen scraper*. Screen scrapers designed for blogs have two main parts: a *mapper* that identifies a unique URL for every post on the blog, and a *parser* that extracts relevant fields from each post. From a programming perspective, the mapper must understand the blog's routes,

[8] Another possibility would be to collect comments from blog posts. The scope of data collection was already so extensive that I decided not to attempt to parse any comments themselves.

[9] For administrative reasons, the target sample for blog parsing was limited to bloggers eligible for contact in the third wave of the survey. This reduces our effective sample size somewhat, but should not introduce any bias, since wave 3 blogs were sampled using the same stratification approach as in previous waves.

and the parser must understand the blog's HTML/CSS template.

Writing custom scrapers for thousands of different blog formats would simply not be possible. Fortunately, the majority of blogs follow a handful of common formats. In particular, a large fraction of blogs are served either by Google's Blogger service or by WordPress. Blogger is the most popular blog hosting service, comprising 41.0% of blogs in the sample. For publicly accessible blogs served by Blogger, an API exposes content in a common format, so we can map and parse with them with 100% coverage.

WordPress is also a very popular blog hosting site, making up 26.2% percent of our target sample. For these blogs, all routes follow one of two formats, so we can map posts with perfect accuracy. In addition, within WordPress blogs, most page layouts fall into a small handful of templates, so we can achieve a reasonably high degree of coverage (89.6%) in parsing.

Outside of Blogger and WordPress, things get more difficult. I was able to identify three other route schemas that collectively cover 180 blogs. In addition, to boost the number of bridge cases between survey results and blog posts, I wrote 96 custom mappers for otherwise unmappable blogs of survey respondents. These two sets of mappers bring our collective map rate up to 72.1% percent for all blogs, and 96.5% percent among survey respondents[10]. Table A.3 reports coverage statistics for mapping and parsing various blog formats.

Unfortunately, the templates of these "Other" blogs (and the 10.4% of unparsed Wordpress blogs) were so varied that parsing our target fields was impossible. This imposes a serious penalty on our parse rates, which come out to 89.6% overall and 74.0% among survey respondents. At the end of the day, the effective response rates (i.e. the percent of blog that were successfully mapped *and* parsed) are still reasonably high: 65.2% for all blogs and 71.5% for survey respondents. Table A.3 reports mapping

[10] This map rate excludes 53 blog sites that were abandoned and deactivated between the survey administration in late 2010, and blog parsing conducted in early 2012. If we include the 53 broken sites in the denominator, the effective map rate is 88.1%.

Table A.2: Map and parse rates by blog type

	Blogger	Wordpress	Other	Total
Eligible	2,356	1,506	1,876	5,738
Mapped	2,356	1,506	276	4,138
Parsed	2,356	1,350	0	3,706
Map rate	100.0%	100.0%	14.7%	72.1%
Parse rate	100.0%	89.6%	0.0%	89.6%
Response rate	100.0%	89.6%	0.0%	64.6%

The map rate equals mapped blogs divided by eligible blogs. The parse rate equals parsed blogs divided by mapped blogs. The responses rate equals parsed blogs divided by eligible blogs.

Table A.3: Map and parse rates among survey respondents' blogs

	All	Available
Eligible	603	550
Mapped	531	531
Parsed	393	393
Map rate	88.1%	96.5%
Parse rate	74.0%	74.0%
Response rate	65.2%	71.5%

53 blogs were deleted between the time the survery was administered, and the final blog post crawl was conducted. These blogs are excluded from the "available" column.

and parsing statistics for the blogs of survey respondents.

In addition, we can salvage content from the 276 blogs that were mapped but not parsed using a technique I call *diff-based parsing*. It works as follows. First we create a full inventory of blog posts using the appropriate mapper. Second, we download the full HTML content of each post page. Next, we clean each post to remove non-HTML elements (e.g. CSS, Javascript) and nonstandard HTML attributes, but not the essential HTML structure of the page. As a last preprocessing step, we select a small reference set of blog posts at random.

With these preliminary steps complete, we apply two steps to each post within each blog. First, we apply the diff operator against each post in the reference set to generate the list of edits required to change the reference post into the target post. Second, we re-create the unique elements of the target post by taking the union of edits from diffs against the full reference set.

This approach effectively strips out any formatting and page elements that are held in common among blog posts, leaving only the unique elements of each post. This will typically include the title, date, labels, body text, and comments of the post. Diff-based parsing doesn't allow us to associate these bits of content with specific fields, but it does allow us to isolate the raw content of each post from the overall format of the blog. The main drawback to diff-based parsing is that it doesn't capture dates or disassociate comments and body text. Still, it helps us recover data that would have otherwise been completely lost from analysis.

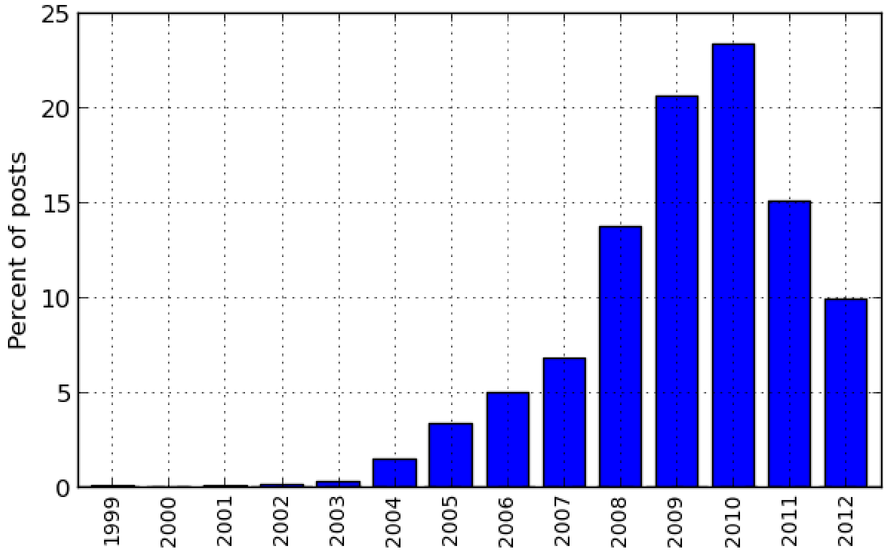
A.4.2 Panel characteristics

In terms of total sample size, the result is staggering, especially when compared to the sample sizes of previous studies of political activism. On average, the blogs in the sample had 1,891 posts each, yielding a total corpus of 7,825,649 mapped posts and 6,831,429 parsed posts.

Figure A.2 shows the percentage of blog posts by year. As shown in the graph, 2010 was the year with the single most blog posts in our panel. Post volume falls off on either side. It's tempting to jump to conclusions about the intensity of the 2010 election cycle or the popularity of political blogging. However, the peak in 2010 is actually caused by our sampling strategy: the sampling universe was constructed in 2010 and the peak in 2010 is a natural reflection of that fact. A significant fraction of active bloggers in 2010 had stopped maintaining their blogs by 2012. Since more recent blogs are not included in the sample, the data show a dropoff in 2011 and 2012.

Figure A.3 illustrates the richness of this data set, using a cross-section of blog posts around the 2008 Presidential elections. For each day leading up to and following the election, I searched all blog posts for keywords corresponding to the four Democratic frontrunners (Obama, Clinton, Edwards, and Biden). The graph shows

Figure A.2: Blog post volume over time



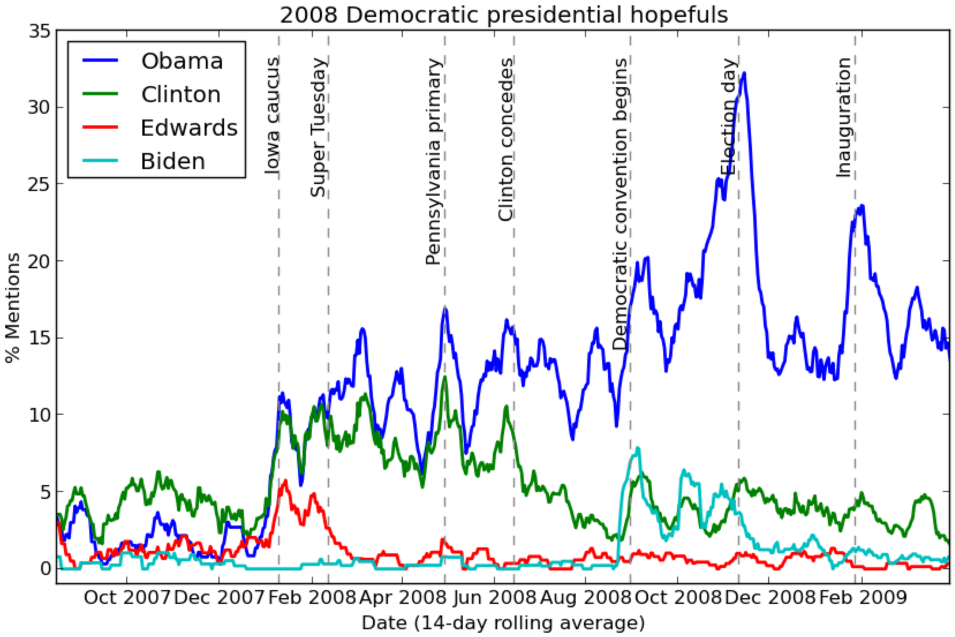
each candidate's day-over-day trend as a separate line, with captions corresponding to the dates of important campaign events, such as the Iowa caucus, super Tuesday, Clinton's concession speech, the Democratic convention, and election day itself. Even this very basic analysis reveals trends in attention to specific candidates, as well as bursts of interest to the campagin in general. On election day, fully one in three blog posts mentioned Barack Obama!

This example is intended as a simple illustration of the richness and scale of the blog post panel. Chapter III introduces more systematic methods for extracting valid measures from this vast corpus of content.

A.5 What makes this research design special?

As I've talked to people about this dissertation, the label that seems to stick most often is "Big Data." As shorthand, I accept the label. But we should note that from the persepctive of scientific inference, "big" matters far less than "carefully structured." At best, more observations allow us to tease out patterns and variation that would

Figure A.3: Mentions of Democratic Presidential hopefuls during the 2008 campaign



have been too noisy to see otherwise. At worst, more observations let us achieve the same results after a lot more work. The key feature of this research design is not its *scale*, but its *structure*: combining a sampling frame, survey, and content analysis. Following are four characteristics that make this combination unique and valuable.

First, the design allows us to *compare and contrast populations*. In sampling and surveying bloggers, my goal has been to construct a representative sample of political bloggers and enable comparisons with the population of U.S. residents in general. Several previous studies have mined large amounts of web data to make social inferences. However, lacking a well-defined sampling universe, they are unable to make claims about the composition of online activism as a whole. In contrast, my methodology produces results that generalize to the whole political blogosphere. For the first time, we can get a clear picture of how bloggers—including little-read bloggers in the tails of the distribution—compare to the U.S. population in general.

Second, the design enables us to *link different aspects of behavior.* Given the option to directly analyze the content of political blogs on a massive scale, it might be tempting to forgo the expense and difficulty of contacting bloggers with a survey. But such a research design would dramatically limit our ability to connect old theories with new data. Many of our best existing social theories are expressed primarily in terms of the best previously existing data sources: survey responses. They have little to say directly about the kinds of behaviors that are newly observable in blogs. Only by observing both kinds of data, side by side, can we relate well-established principles from opinion research and participation theory to the rich new data available online.

Third, this research design allows us to *observe behavior over time.* Lacking the ability to conduct experiments, observations over time are one of the best ways to imbue observational data with causal significance. For that reason, panels are an especially fruitful data structure. In many ways, the real advantage to collecting a very large panel of blog posts is our ability to observe differences in behaviors at fine-grained time scales—days, rather than months or election cycles.

Finally, this research design *draws on two very different methodological traditions.* In order to collect these data, I was forced to mix and match techniques from social and computer science. Broadly speaking, social science is better versed in issues of sampling and measurement, while computer science is more adept at problems in data processing and scalability. As we will see, methods from these two disciplines turn out to be much more powerful together than they are individually. In my opinion, demonstrating the combined analytical power of these tool sets is one of the most important contributions of this dissertation.

APPENDIX B

Survey administration

Because this was the first time a representative survey had been conducted among political bloggers, many of the mechanisms for administering such a survey were untested. Accordingly, a series of waves gave me opportunities to test and improve my methods, while expanding the effective sample size. This appendix provides details about the goals and timing of each of the survey waves (B.1), details about contacting (B.2), the full text of contact emails (B.3), and the survey FAQ linked from the emails (B.4). Appendix C contains the full survey instrument.

B.1 Survey waves

Wave one was conducted in October 2010, just prior to the midterm election. This was essentially a pilot wave, meant to test my apparatus for contacting bloggers, as well as the survey instrument itself. This wave was based on a small sample of 172 bloggers, of whom 65 were contacted, and 26 responded.

Wave two was conducted about six weeks later, in November. This wave expanded the respondent sample to 173, providing enough statistical leverage to conduct tests on effects of incentives and contacting methods. Finding no important effects from either

Table B.1: Changes in survey administration by wave

Wave	Date	Incentives	Method	Sample	Responses
Wave 1	October 2010	Yes	Email	172	26
Wave 2	December 2010	Yes	Email and other	961	147
Wave 3	May 2011	No	Email	5,738	603

treatment[1], I decided to economize by removing incentives and non-email contacting from the last wave of the survey.

Wave three was the largest of the waves, with an eligible sample of 5,738 and 603 respondents[2]. Fielded in May 2011, the questionnaire was modified slightly so that the battery of political knowledge questions would reflect changes in office-holding and the composition of the House and Senate.

Except for non-email contacting, incentives, and the few questions already mentioned, all aspects of survey administration were identical between waves. Table B.1 shows adjustments in survey administration by wave.

B.2 Contacting

Gathering contact information for bloggers was perhaps the most difficult part of conducting the survey. During the filtering phase, the research team searched each blog thoroughly for an email address to contact the blogger. Some blogs are co-authored by multiple bloggers. In these cases, we used a Kish grid to randomize selection of one blogger for email contacting. This process is almost identical to the process for selecting members within households in door-to-door contacting.

More commonly, we confronted the opposite problem: no email address at all. Only about one in three political blogs includes an email address to contact the author(s). This is similar to the problem of cellphone-only households that is currently disrupting

[1] In pairwise tests against every variable in the survey, incentives and contacting methods showed "significant effects" in a handful of cases, but no more than would be predicted by the natural rate of false positives.

[2] This figure is pessimistic, excluding 291 partially completed surveys.

telephone surveys.

Arbitrarily excluding two-thirds of the sample would present a serious threat to the validity of the study. To alleviate this problem, I adopted the following system in the second wave of the survey: prior to final screening, I randomly assigned each blog to one of two contacting conditions: "email-only" or "by any means." Eighty percent of blogs were assigned to the email-only condition. Within this set, bloggers with email addresses were invited to participate in the study; other were left out of the survey, but retained in the sampling frame.

The other twenty percent of blogs were contacted by any means available. For convenience, we used email when possible. When no email address was available, we next looked to send the invitation using a "drop box" for posting comments privately to the blogger. When no drop box was available, we posted the invitation as a comment to the most recent blog post, even if the comment was publicly visible. A few of the sampled blogs did not even allow comments and could therefore not be contacted.

This protocol economizes on effort spent on contacting and sample recruitment without excluding hard-to-contact bloggers from the sample entirely. Accordingly, in the second wave of the survey, 384 bloggers received survey invitations by email, and 79 bloggers received invitations by drop box or comment. Only a handful of bloggers could not be contacted at all.

As shown in Table B.2, this approach resulted in a moderate improvement in effective response rates. However, after analyzing results from the second wave, I saw no important differences between respondents contacted by email and by other means[3]. Given that non-email contacting was far more expensive than contacting by email and didn't seem to have much impact on the substantive results, I decided to forgo non-email contacting in the final wave of the survey.

In addition, for email invitations in the second wave, I included randomized

[3] These tests were based on sample sizes with admittedly low statistical power. Still, the results suggest that any biases induced by excluding bloggers without emails addresses are likely to be small.

Table B.2: Eligibility, contacting, and responses by survey wave

	Wave 1	Wave 2 Email-only	Wave 2 Any means	Wave 3	Total
Eligible	172	777	184	5,738	6,871
Contacted	65	256	142	2,609	3,072
Responded	26	104	43	603	776
Contact rate	37.8	32.9	77.2	45.5	44.7
Cooperation rate	40.0	40.6	30.3	23.1	25.3
Response rate	15.1	13.4	23.4	10.5	11.3

incentives. At the time of the first email contact, bloggers were randomly assigned offers of no incentive, $10, or $20, as “a small thank-you for participating.” These offers were persistent, with reminders in subsequent emails and the instructions to the survey itself. At the end of the survey, respondents in either of the two incentive conditions were able to fill out a form to receive an Amazon gift card by email. As with non-email contacting, incentives proved to have little or no effect on survey responses, so I did not use incentives in the final wave of the survey.

My contacting strategy for drop boxes and comments was similar to email contacting, with two exceptions. First, I sent only one follow-up message—since these bloggers had decided not to provide contact information, I felt that sending two unsolicited messages was stretching the no-spam etiquette of the Internet far enough. Second, Qualtrics’ system of authentication made it difficult to match non-emailed links with incentives. Therefore, I only offered incentives to respondents contacted by email.

All told, I was able to gather contact information for 44.7% of sampled bloggers, and achieve a cooperation rate of 25.3% , with an effective response rate of 11.3% . These response rates are lower than we might like, but are still quite strong for Internet research. It seems likely that much of the non-response is being driven by inactive blogs and email accounts. Moreover, tests based on randomly assigned incentives and modes of contact give us some evidence that sample bias due to non-response is likely to be small. Finally, the availability of full blog content opens the possibility for unusually good sample weighting to adjust for potential issues in the sampling process. I describe methods for taking advantage of this useful feature of blog data in Chapter II.

In the end, this contacting strategy was successful, yielding a final sample of 776 respondents, drawn with known probability from the general population of active, English-language U.S. political bloggers.

B.3 Contact emails

A note on [used_name]:

If we know a blogger's first name, then used_name is the same as first_name: "Ted" Otherwise, if we know a blogger's psuedonym, then used_name is the pseudonym: "M1dn1ght Bl0gger" Finally, if we don't have either, then we just use the name of the blog: "Random Progressivist Thoughts"

B.3.1 Initial contact

`[Used_Name] -`

We are writing to ask for your help in a survey of bloggers being conducted by the University of Michigan. This study looks at how the blogging community thinks about and responds to important issues and current events. As `[Authors_Plural]` of `[Blog_Name]`, you are invited to visit the link below and complete the survey. It will only take about 20 minutes.

We hope you will find the survey interesting and enjoyable. Simply click on the link below, or cut and paste the entire URL into your browser to access the survey:

`[Survey_Link]`

We appreciate your participation – it is only by hearing from nearly everybody in the sample that we can be sure the results are truly representative. Your responses will be kept strictly confidential, and we will be happy to share our results with you once the study is complete. [Incentive_Text]

Thanks in advance for your willingness to share your ideas and experiences.

Abe Gong and Nancy Burns
University of Michigan

P.S. If you have questions or comments about this study, please direct them to Abe Gong (agong@umich.edu). Thank you!

B.3.2 First reminder

[Used_Name] -

A few days ago, we sent you a link to a survey seeking your opinions as a blogger about issues and events in society today. If you have already completed the survey, please accept our sincere thanks. If not, please do so today by following the link below.

[Survey_Link]

The survey is only 20 minutes long, and all your responses will be completely confidential. [Incentive_Text_2] Our goal is to assemble a clear picture of ideas, opinions, and demographics in the blogosphere. We are grateful for your participation, because every response will make the final results more accurate.

Abe Gong and Nancy Burns
University of Michigan

P.S. If by some chance we have made a mistake and you are not [Used_Name] [Last_Name] from [Blog_Url], please answer at least the first three questions of the survey to help us clear up the misundertanding. Many thanks!

B.3.3 Second reminder

[Used_Name] -

About a week ago, we sent you a link asking for your help on a survey of bloggers. This email is a quick reminder to complete the survey, before it closes next Monday, Dec. 20. We know you are busy, and appreciate your participation in this study of opinions and habits in the blogosphere.

[Survey_Link]

Thank you and enjoy the survey!

Abe Gong and Nancy Burns

University of Michigan

P.S. A comment on our survey procedures. Although we use your name and email address to verify survey completion, those fields are deleted and replaced with a random identification code number once you take the survey. Consequently, all survey responses are completely confidential. We are happy to share aggregate results, but we will never release results or data that could be used to identify individual survey respondents. Protecting the confidentiality of people's answers is very important to us, as well as the University.

B.3.4 Thank you

[Used_Name] -

Thanks again for responding to our survey of ideas and opinions in the blogosphere. We appreciate your participation – the only way to learn what bloggers think is to ask bloggers themselves!

This study will be in the field until the end of April next year. In the meantime, if you have questions or comments, please let us know. We're interested in improving the survey for future respondents, and welcome your feedback.

Please remember to not talk or blog about this study until after April. We want to make sure that the survey experience is consistent for everyone. Once the study is complete, we will be happy to share the results, and you will be completely free to talk about it.

If you have further questions or comments about this study, please contact Abe Gong (agong@umich.edu). Thank you!

Abe Gong and Nancy Burns
University of Michigan

B.4 Frequently asked questions

Who are you?

I'm a graduate student studying political science at the University of Michigan. Nancy Burns is my dissertation advisor. You can find my bio and webpage through the department website.

Why are you doing this project?

This survey is part of a dissertation project seeking to understanding how people discuss current events and social issues online. There is no commercial or political agenda behind the project. I'm just trying to get a clear picture of who participates in the blogosphere.

How did you find my blog?

I found sites for this survey as part of a "snowball sample," following links between web sites to identify bloggers who post on topics related to current events. The overall sample will include about 700 bloggers, with the goal of getting a balanced picture of the views and ideas in the blogosphere.

Does this project have IRB approval?

The survey protocol has been reviewed and approved by the University of Michigan IRB.

How can I know this isn't some kind of scam?

You can find me on the list of current University of Michigan political science grad students at http://polisci.lsa.umich.edu/grad/currentGrads.html. You can also email me at agong [at] umich [dot] edu, and I'll respond as soon I can.

How long will the survey take to complete?

About 20 minutes. I know that you're busy and 20 minutes is a lot to ask. I appreciate you time and opinions.

Are there any incentives to complete this survey?

Sadly, as a graduate student working on a grad student budget, I can't pay people to participate in this survey. If it helps, I'm not making any money from the project either. Once the survey is complete, I'll be happy to share (anonymized) results.

In the meantime, this survey is a chance to make sure that your views and opinions on blogging and current events are represented. Our goal is to get a clear and balanced picture of who participates in the blogosphere – it won't be complete unless almost almost everyone invited participates.

I didn't receive an invitation to the survey. Can I still participate?

Yes, but not right now. Send me an email agong [at] umich [dot] edu, and I'll let you know about future surveys.

How will my confidentiality be protected?

Although we use your name and email address to verify survey completion, those fields are deleted and replaced with a random identification code number once you take the survey. Consequently, all survey responses are completely confidential. We are happy to share aggregate results, but we will never release results or data that could be used to identify individual survey respondents. Protecting the confidentiality of people's answers is very important to us, as well as the University.

Some of the questions are a little personal. What if I'm not comfortable answering them?

Most of the questions come from mainstream surveys like Gallup, Pew, and the General Social Survey. That said, we respect your privacy, so if there are questions you're not comfortable answering, you're welcome to skip them and go on.

There was a mistake on my name, or the name of my blog.

I'm very sorry if we got your name wrong. A team of research assistants was responsible for collecting names and contact information for this survey, and they make some mistakes in the process.

My blog doesn't get a lot of traffic. Do you still want me to respond?

We're trying to get a balanced picture of the views and ideas in the blogosphere, and that includes small blogs as well as big ones. It won't be a balanced picture if we only hear back from the people with a lot of readers.

I'm not from the U.S. Should I still respond to the survey?

The survey definitely has a U.S. slant, mainly because so many bloggers are American. That said, we'd appreciate your responses on the questions that are relevant. You're welcome to skip the others.

I'm interested in the survey results. How can I learn more?

The survey will close in mid-June 2011. I'll be working on analyzing the responses throughout the summer and fall, publishing results along the way. If you have questions, comments, or other feedback, please contact me at agong [at] umich [dot] edu. My goal is to learn what theories of participation and communication can teach us about the blogosphere, and vice versa. I welcome suggestions and constructive criticism.

APPENDIX C

Survey instrument

Introduction

Thank you for your participation in this survey of bloggers, sponsored by the University of Michigan!

Your blog, ${e://Field/blog_name} at ${e://Field/blog_url}, has been selected as part of a small, representative sample of blogs discussing important issues and current events in society. In this survey you will be asked about your blogging habits and reasons for blogging, as well as your ideas and opinions on things in the news, current events, and so on.

- Your **participation is voluntary**. You are free not to answer any question or to withdraw from the study at any time. This survey should take about 15 minutes to complete.
- Your **responses will be completely confidential** -- no identifying information will be made public. Data from this survey and some corresponding statistics from your blog will be stored at a secured location and retained indefinitely.
- Once all survey responses are in, we will be happy to share results with you. Until then, **we ask you to not discuss or blog about the survey**. If you have any questions about this study, please contact Abe Gong (agong@umich.edu).

Before we continue, we need to verify some information about you and your blog.

Are you the author (or one of the authors) of the blog ${e://Field/blog_name}?

○ Yes, I am an author of ${e://Field/blog_name}.

○ No, I am not an author on any blog.

○ No, I am not an author of ${e://Field/blog_name}, but I am an author of a different blog:

______

Are you at least 18 years old?

○ Yes

○ No

Which best describes your citizenship status in the United States?

○ I was born and currently live in the U.S.

○ I am a U.S. citizen currently living outside the country.

○ I am an immigrant to the U.S. and a naturalized citizen.

○ I am an immigrant to the U.S. but not a citizen.

○ I do not live in the U.S. and am not a citizen.

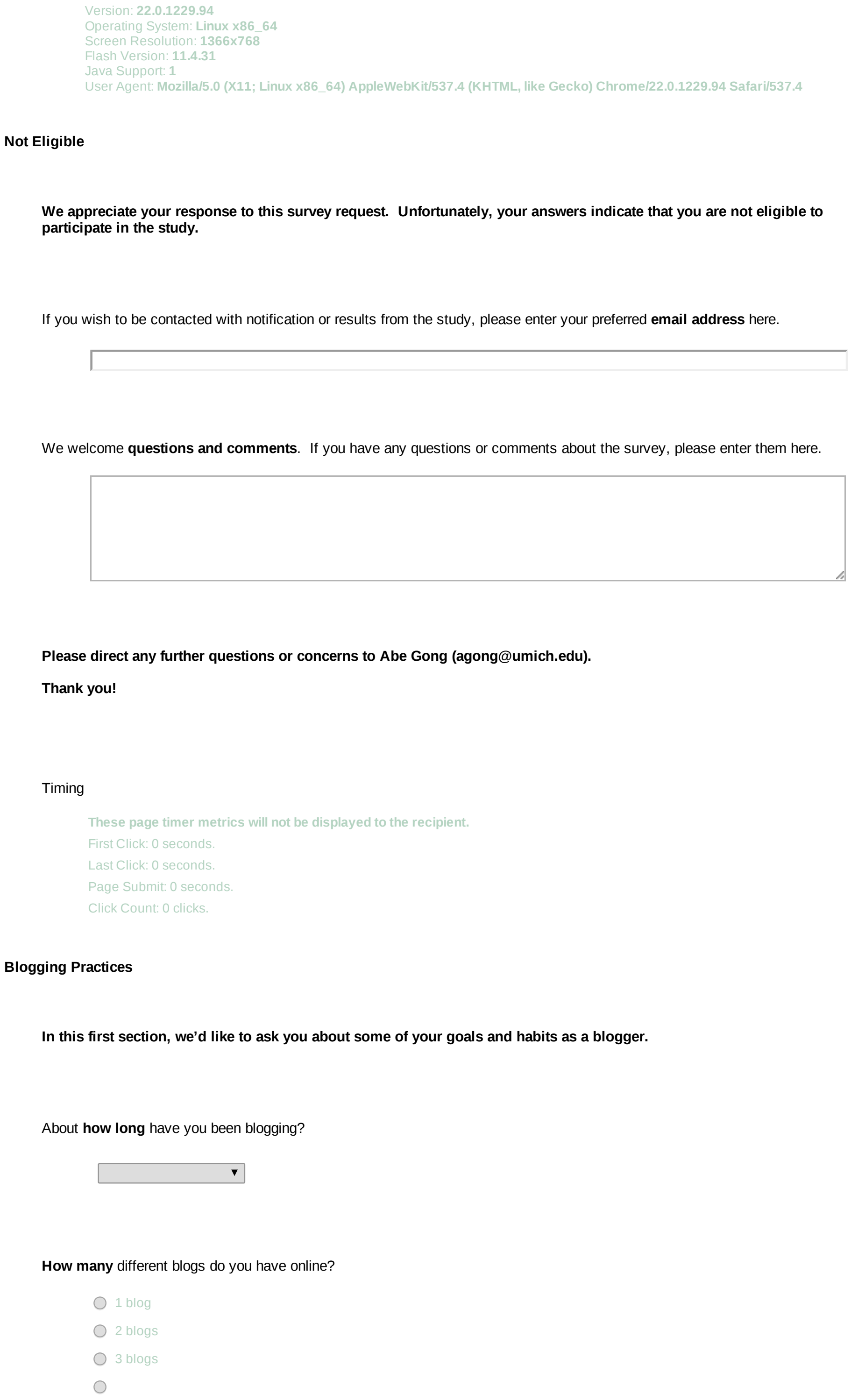
To begin the survey, please click the "Next >>" button at the bottom of the page. By doing so, you indicate that you voluntarily consent to participate in this study under the terms given above.

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

Browser Meta Info

This question will not be displayed to the recipient.
Browser: **Chrome**



Version: **22.0.1229.94**
Operating System: **Linux x86_64**
Screen Resolution: **1366x768**
Flash Version: **11.4.31**
Java Support: **1**
User Agent: **Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.4 (KHTML, like Gecko) Chrome/22.0.1229.94 Safari/537.4**

Not Eligible

We appreciate your response to this survey request. Unfortunately, your answers indicate that you are not eligible to participate in the study.

If you wish to be contacted with notification or results from the study, please enter your preferred **email address** here.

We welcome **questions and comments**. If you have any questions or comments about the survey, please enter them here.

Please direct any further questions or concerns to Abe Gong (agong@umich.edu).

Thank you!

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

Blogging Practices

In this first section, we'd like to ask you about some of your goals and habits as a blogger.

About **how long** have you been blogging?

▼

How many different blogs do you have online?

○ 1 blog
○ 2 blogs
○ 3 blogs
○

4 blogs

○ 5 or more blogs

Which best describes your **involvement** with blogging?

○ I blog for fun. I do not make, or ever plan to make, any money on my blog.

○ Right now I blog for fun. I would like to make money on my blog some day.

○ I use my blog to supplement my income, but I don't consider it my full time job.

○ I blog full-time for a company or organization.

○ I blog full-time for my own company or organization.

Here are some **reasons** people give for blogging. How important are they for you?

For each statement, please check the response that best describes you.

	Very important	Somewhat important	Not too important	Not at all important
I blog to meet and connect with like-minded people.	○	○	○	○
I blog to share practical knowledge or skills with others.	○	○	○	○
I blog to make money or supplement my income.	○	○	○	○
I blog to share my experiences with others.	○	○	○	○
I blog to encourage social change on issues I care about.	○	○	○	○
I blog to keep friends and family updated on my life.	○	○	○	○
I blog to express myself creatively.	○	○	○	○
I blog to speak my mind on areas of interest.	○	○	○	○
I blog to promote my business, product, or resume.	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

For the questions on this page, we'd like to focus on your blog ${e://Field/blog_name}. Please think only about this blog for these questions.

What would you say is the **main topic** of your blog, ${e://Field/blog_name}?

For the questions on this page, we'd like to focus on your blog ${q://QID2/ChoiceTextEntryValue/3}. Please think only about this blog for these questions.

What would you say is the **main topic** of your blog ${q://QID2/ChoiceTextEntryValue/3}?

□

What **other important topics** does this blog cover?

□

How many entries did you post on your blog last week?

▼

How many days did you work on your blog last week?

0 days	1	2	3	4	5	6	7 days
○	○	○	○	○	○	○	○

Approximately **how many hours** did you spend working on your blog per day last week?

▼

Including yourself, **how many authors** does ${e://Field/blog_name} have?

(By authors, we mean regular contributors who generate content for the blog. Please don't count guest authors who contribute only by invitation.)

○ I am the only author.

○ 2 authors

○ 3 authors

○ 4 authors

○ 5 or more authors

Including yourself, **how many authors** does your blog ${q://QID2/ChoiceTextEntryValue/3} have?

(By authors, we mean regular contributors who generate content for the blog. Please don't count guest authors who contribute only by invitation.)

○ I am the only author.

○ 2 authors

○ 3 authors

○ 4 authors

○ 5 or more authors

About how many **page views** did your blog receive last month?

□

About how many **unique visitors** did your blog receive last month?

□

In the last year, has your blog **received attention** from or been **mentioned by** any of the following?
(Please check all that apply)

□ Public officials, politicians, or political campaigns

□ The traditional news media

□ Other bloggers

□ Local community members

□ Colleagues, coworkers or bosses

□ Family members

□ Companies

Next, we are going to ask you to choose which of two statements comes closer to your own opinion. You might agree to some extent with both, but we want to know **which one is closer** to your own views.

I mainly try to inform people using my blog.	○	○	○	○	○	○	○	I try to get my readers to take action.
I get inspiration for my blog posts mainly from everyday life.	○	○	○	○	○	○	○	I get inspiration for my blog posts mainly from the news and current events.
I blog mostly for myself.	○	○	○	○	○	○	○	I blog mostly for my audience.
I try to be objective and unbiased in my blog.	○	○	○	○	○	○	○	My blog is a place where I express my views and opinions.
I blog mostly about things that are meaningful to me, personally.	○	○	○	○	○	○	○	Attracting and retaining an audience for my blog is important to me.
For the most part, my readers and I share a common world view.	○	○	○	○	○	○	○	Readers of my blog and I often have very different opinions.
When I blog, I try to be engaging and entertaining.	○	○	○	○	○	○	○	When I blog, I try to communicate the facts clearly.

About what percent of the content on your blog is about politics?

▼

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

We're interested in learning about people who **make money** through their blogs.

Still thinking about your blog ${e://Field/blog_name}, please check the statements that apply to you.

□ I make money on my blog through advertising.

□ I make money on my blog through reader contributions, like a “tip jar.”

□ I make money on my blog through premium content that is only accessibly by paying a fee.

□ I make money on my blog by selling items on my site.

□ I am paid to write this blog.

□ I write for this blog as part of another job.

□ I do not make any money on my blog.

Still thinking about your blog ${q://QID2/ChoiceTextEntryValue/3}, please check the statements that apply to you.

□ I make money on my blog through advertising.

□ I make money on my blog through reader contributions, like a “tip jar.”

□ I make money on my blog through premium content that is only accessibly by paying a fee.

□ I make money on my blog by selling items on my site.

□ I am paid to write this blog.

□ I write for this blog as part of another job.

□ I do not make any money on my blog.

In the last year, about **how much advertising revenue** did you make through your blog?

▼

In the last year, about **how much TOTAL revenue** did you make through your blog?

▼

In the last year, about **how much money did you spend** running your blog?

(Common costs of running a blog include site development and maintainence, staffing, marketing and advertising, hosting fees, etc.)

▼

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

News and Media Consumption

Next, we’d like to ask some questions about your Internet use.

About **how long** have you been an Internet user?

▼

How many days did you go online in the last week?

0 days	1	2	3	4	5	6	7 days
○	○	○	○	○	○	○	○

Approximately **how many hours per day** did you spend online last week?

▼

What **devices** do you regularly use to access the Internet?

- □ Desktop computers
- □ Laptop computers
- □ Mobile devices, such as iPhones or Blackberries
- □ Other

How many days did you read other people's blogs last week?

0 days	1	2	3	4	5	6	7 days
○	○	○	○	○	○	○	○

Approximately **how many hours per day** did you spend reading other people's blogs last week?

▼

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

We'd like to know about sources where you get news and information.

How often do you get **news about politics** from each of the following sources?

	Never	Less than Once a Month	Once a Month	2-3 Times a Month	Once a Week	2-3 Times a Week	Every day
Radio	○	○	○	○	○	○	○
Paper newspapers	○	○	○	○	○	○	○
Magazines	○	○	○	○	○	○	○
Internet blogs	○	○	○	○	○	○	○
Television	○	○	○	○	○	○	○
Internet news sites	○	○	○	○	○	○	○

How often do you watch each of the following on television or the Internet?

	Never	Occasionally	A few times a week	Every night or almost every night
National news programs	○	○	○	○
Local news programs	○	○	○	○
Cable news networks	○	○	○	○
The Rachel Maddow Show	○	○	○	○
The Daily Show with Jon Stewart	○	○	○	○
Hardball with Chris Matthews	○	○	○	○
The O'Reilly Factor	○	○	○	○
The Newshour with Jim Lehrer	○	○	○	○
Countdown with Keith Olbermann	○	○	○	○
Hannity	○	○	○	○

We'd also to like know a little about how you use your time in general.

In the last week, about how many **hours per day** did you spend ...

... **on necessary work for your home and family** such as cooking, cleaning, taking care of children or relatives, shopping, house and yard chores,etc.? ______

... **on gainful employment**, including commuting and work that you take home? ______

... **studying for a degree or enrolled in courses**? ______

... **sleeping**? ______

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Political Attitudes

In this next section, we'd like to hear about some of your opinions about politics.

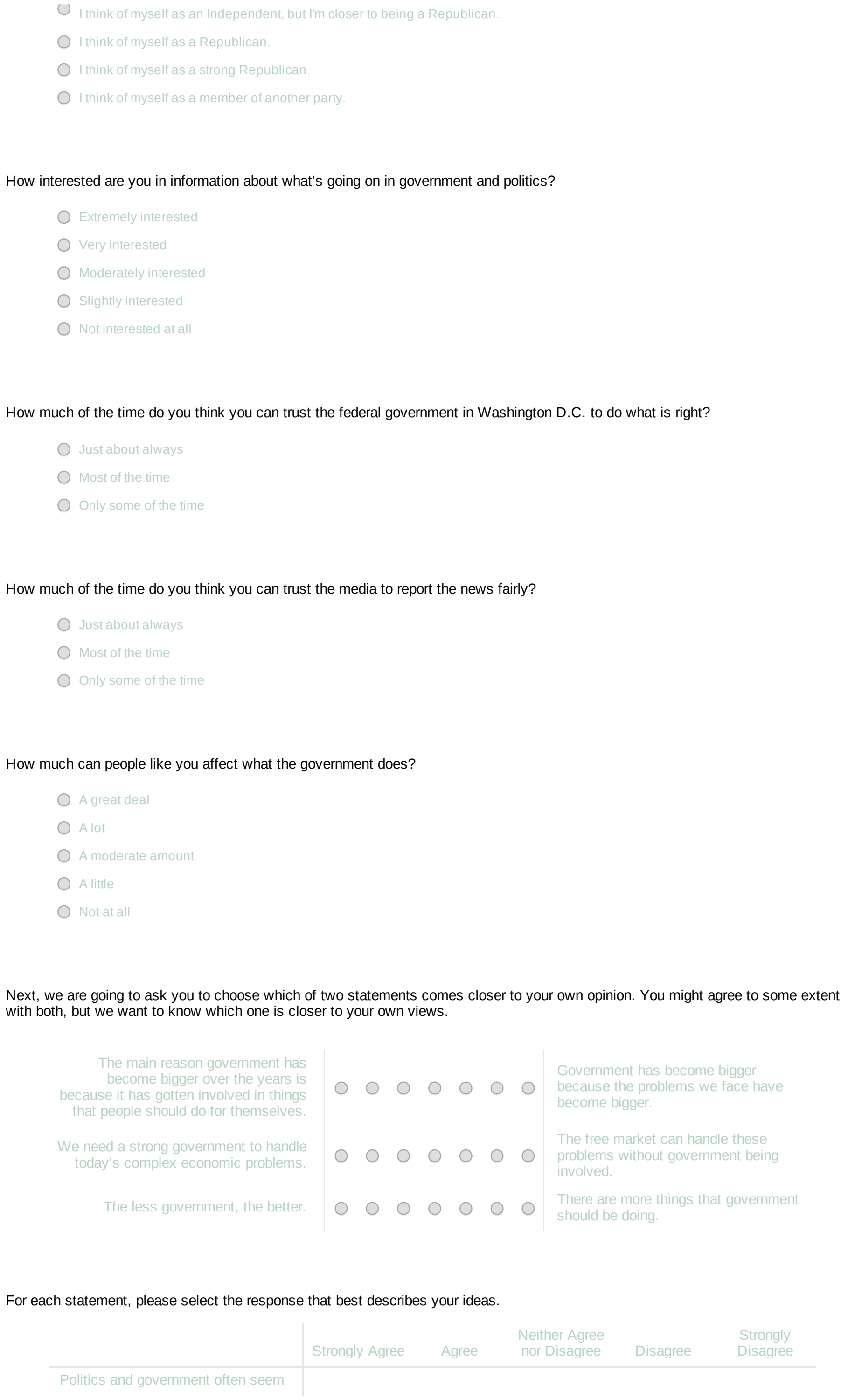
Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?

○ I think of myself as a strong Democrat.

○ I think of myself as a Democrat.

○ I think of myself as an Independent, but I'm closer to being a Democrat.

○ I think of myself as an Independent.



○ I think of myself as an Independent, but I'm closer to being a Republican.

○ I think of myself as a Republican.

○ I think of myself as a strong Republican.

○ I think of myself as a member of another party.

How interested are you in information about what's going on in government and politics?

○ Extremely interested

○ Very interested

○ Moderately interested

○ Slightly interested

○ Not interested at all

How much of the time do you think you can trust the federal government in Washington D.C. to do what is right?

○ Just about always

○ Most of the time

○ Only some of the time

How much of the time do you think you can trust the media to report the news fairly?

○ Just about always

○ Most of the time

○ Only some of the time

How much can people like you affect what the government does?

○ A great deal

○ A lot

○ A moderate amount

○ A little

○ Not at all

Next, we are going to ask you to choose which of two statements comes closer to your own opinion. You might agree to some extent with both, but we want to know which one is closer to your own views.

The main reason government has become bigger over the years is because it has gotten involved in things that people should do for themselves.	○	○	○	○	○	○	○	Government has become bigger because the problems we face have become bigger.
We need a strong government to handle today's complex economic problems.	○	○	○	○	○	○	○	The free market can handle these problems without government being involved.
The less government, the better.	○	○	○	○	○	○	○	There are more things that government should be doing.

For each statement, please select the response that best describes your ideas.

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Politics and government often seem					

so complicated that I can't really understand what's going on.	○	○	○	○	○
I feel that I have a pretty good understanding of the important political issues facing our country.	○	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

For the next few questions, we'd like to get your feelings about some groups in American society.

How close do you feel to each of these groups in terms of ideas and interests? Select one answer for each row in the the grid.

	Extremely close								Not close at all		
	0	10	20	30	40	50	60	70	80	90	100
Hispanics or Latinos											
Immigrants											
Men											
Whites											
Conservatives											
Evangelical Christians											
Gay men and lesbians, that is, homosexuals											
Business owners											
The Tea Party movement											
Feminists											
Labor unions											
Blacks											

Liberals											
Environmentalists											
Women											
Asians											

Do you strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, or strongly disagree with these statements?

	Strongly agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
This country would be better off if there were more emphasis on traditional family ties.	○	○	○	○	○
The world is always changing and we should accommodate our view of moral behavior to those changes.	○	○	○	○	○
Discrimination against women is no longer a problem in the United States.	○	○	○	○	○
Society has reached the point where women and men have equal opportunities for achievement.	○	○	○	○	○
Irish, Italians, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.	○	○	○	○	○
Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.	○	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Reasons for Blogging about Politics

Earlier you mentioned that you sometimes blog about politics.

Here are some statements about **reasons** a person might blog about politics. How important are they in your decision to keep up your blog? For each statement, please check the response that best describes you.

When I blog about politics, I do it…

	Very important	Somewhat important	Not too important	Not at all important
for the chance to work with people who share my ideals.	○	○	○	○
for the chance to influence government policy.	○	○	○	○
for the chance to further my job or career.	○	○	○	○
because I find it exciting.	○	○	○	○
because it's my duty as a citizen.	○	○	○	○

for the chance to meet important and influential people.	○	○	○	○
to learn about politics and government.	○	○	○	○

When I blog about politics, I do it…

	Very important	Somewhat important	Not too important	Not at all important
for the chance to make the community or nation a better place to live.	○	○	○	○
for the chance to further the goals of my party.	○	○	○	○
for the chance for recognition from people I respect.	○	○	○	○
because I might want to get a job with government some day.	○	○	○	○
because I did not want to say no to someone who asked.	○	○	○	○
because I might want to run for office some day.	○	○	○	○
for the chance to be with people I enjoy.	○	○	○	○
because I might want to get help from an official on a personal or family problem.	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Turnout and Vote Choice

Thank you. Next we'd like to hear about who you supported in recent elections.

Are you currently **registered to vote**? If so, are you registered as being a Republican, a Democrat, or something else?

○ I'm not registered to vote.

○ I'm registered as a Democrat.

○ I'm registered as a Republican.

○ I'm registered under another party.

○ I'm registered to vote, but without a declared party.

We'd like to ask you about the national elections for Congress and other offices held in November of last year.
Did you **vote** in this election?

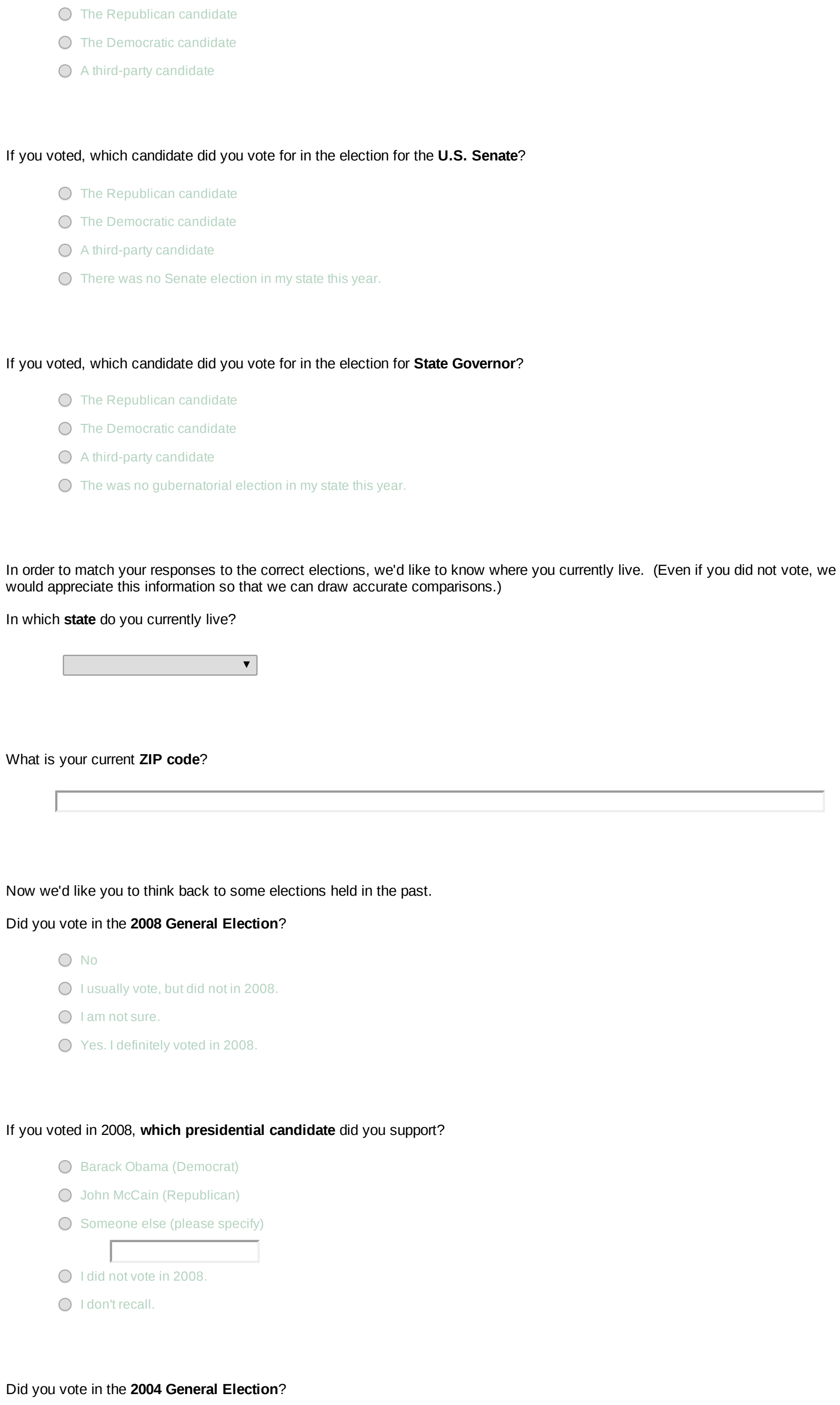
○ No

○ I usually vote, but did not this time.

○ I am not sure.

○ Yes. I definitely voted.

If you voted, which candidate did you vote for in the election for the **U.S. House of Representatives**?



○ The Republican candidate

○ The Democratic candidate

○ A third-party candidate

If you voted, which candidate did you vote for in the election for the **U.S. Senate**?

○ The Republican candidate

○ The Democratic candidate

○ A third-party candidate

○ There was no Senate election in my state this year.

If you voted, which candidate did you vote for in the election for **State Governor**?

○ The Republican candidate

○ The Democratic candidate

○ A third-party candidate

○ The was no gubernatorial election in my state this year.

In order to match your responses to the correct elections, we'd like to know where you currently live. (Even if you did not vote, we would appreciate this information so that we can draw accurate comparisons.)

In which **state** do you currently live?

▼

What is your current **ZIP code**?

Now we'd like you to think back to some elections held in the past.

Did you vote in the **2008 General Election**?

○ No

○ I usually vote, but did not in 2008.

○ I am not sure.

○ Yes. I definitely voted in 2008.

If you voted in 2008, **which presidential candidate** did you support?

○ Barack Obama (Democrat)

○ John McCain (Republican)

○ Someone else (please specify)

○ I did not vote in 2008.

○ I don't recall.

Did you vote in the **2004 General Election**?

○ No

○ I usually vote, but did not in 2004.

○ I am not sure.

○ Yes. I definitely voted in 2004.

If you voted in 2004, **which presidential candidate** did you support?

○ John Kerry (Democrat)

○ George W. Bush (Republican)

○ Someone else (please specify)

□

○ I did not vote in 2004.

○ I don't recall.

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

Community Activity

Next, we'd like to ask about other ways people sometimes get involved in community and government besides blogging and voting.

During the past 12 months, have you **gone to a political speech, march, rally, or demonstration**, or have you not done this in the past 12 months?

○ Yes, I have done this in the past 12 months

○ No, I have not done this in the past 12 months

During the past 12 months, have you **phoned, emailed, written to, or visited a government official** to express your views on a public issue, or have you not done any of these things in the past 12 months?

○ Yes, I have done this in the past 12 months

○ No, I have not done this in the past 12 months

During the past 12 months, have you **worn a campaign button, put a campaign sticker on your car, or placed a sign in your window in front of your house**, or have you not done any of these things in the past 12 months?

○ Yes, I have done this in the past 12 months

○ No, I have not done this in the past 12 months

During the past 12 months, have you **given money to any candidate running for public office, any political party, or any other group that supported or opposed candidates**, or have you not done this in the past 12 months?

○ Yes, I have done this in the past 12 months

○ No, I have not done this in the past 12 months

Thinking back further...

In the past five years, has someone in authority in the following areas asked you to personally **vote for or against certain candidates in an election** for public office?

	No	Yes, once	Yes, twice	Yes, three to five times	Yes, six to ten times	Yes, more than 10 times
At work	○	○	○	○	○	○
At your church or place of worship	○	○	○	○	○	○
In some other organization	○	○	○	○	○	○
From a political campaign	○	○	○	○	○	○

In the past five years, has someone in authority in the following areas asked you to **take some other action on a political issue**, such as sign a petition, write a letter, or get in touch with a public official?

	No	Yes, once	Yes, twice	Yes, three to five times	Yes, six to ten times	Yes, more than 10 times
At work	○	○	○	○	○	○
At your church or place of worship	○	○	○	○	○	○
In some other organization	○	○	○	○	○	○
From a political campaign	○	○	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Continuing on the theme of community involvement...

In the last six months, have you **given a presentation or speech** in the following settings?

	No	Yes, once	Yes, twice	Yes, three to five times	Yes, six to ten times	Yes, more than 10 times
At work	○	○	○	○	○	○
At your church or place of worship	○	○	○	○	○	○
In some other organization	○	○	○	○	○	○

In the last six months, have you **planned or chaired a meeting** in the following settings?

	No	Yes, once	Yes, twice	Yes, three to five times	Yes, six to ten times	Yes, more than 10 times
At work	○	○	○	○	○	○
At your church or place of worship	○	○	○	○	○	○
In some other organization	○	○	○	○	○	○

How often does the subject of politics come up in each of the following situations?

	A lot	Some	Hardly ever	Never	Does not apply
At work	○	○	○	○	○
At your church or place of worship	○	○	○	○	○
In conversations with friends	○	○	○	○	○
In conversations with family	○	○	○	○	○
In conversations on an Internet message board or blog	○	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

Current Events and Issues

We’re almost done. We’d like to get your responses about some current issues and events.

Do you feel things in this country...

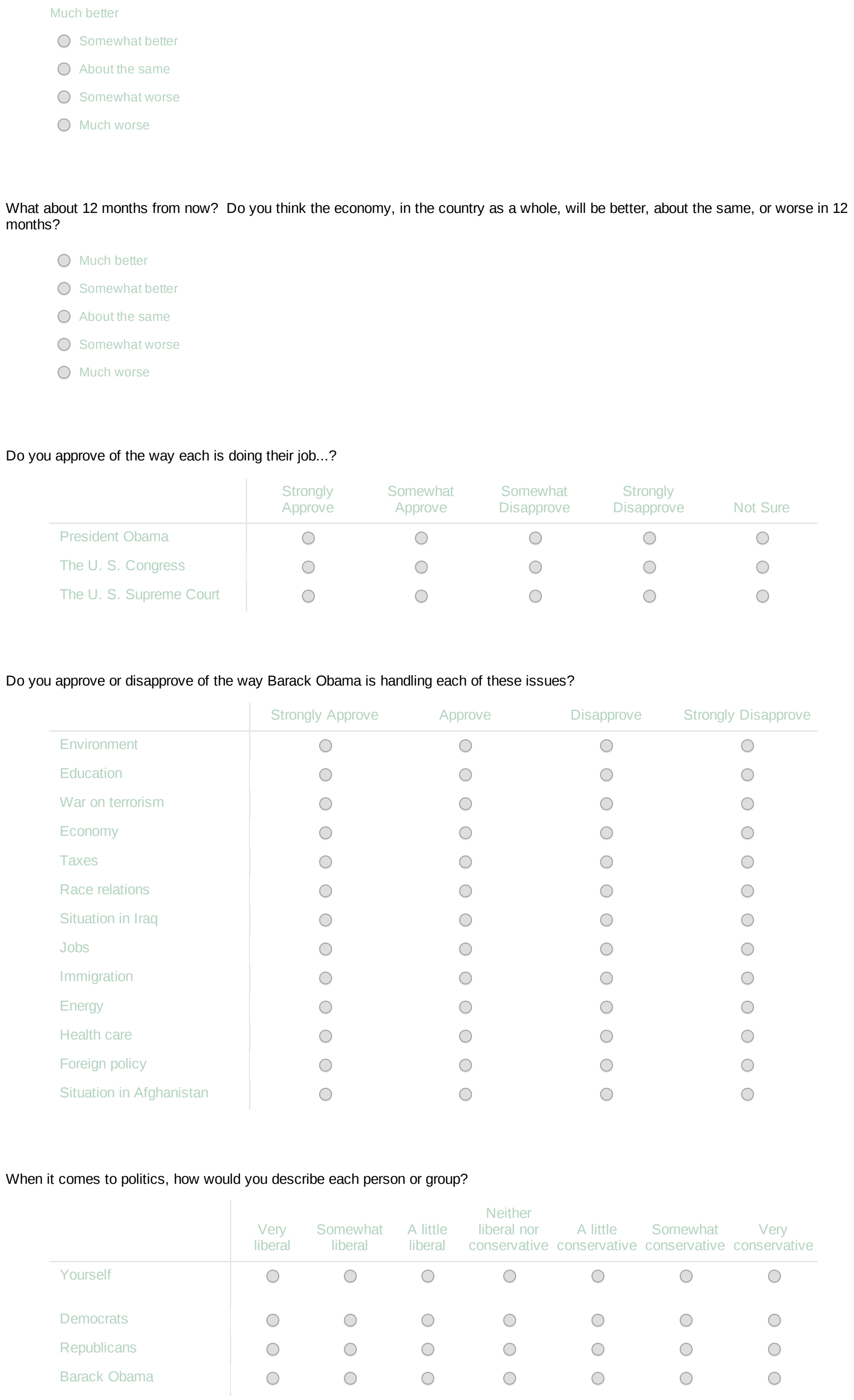
○ Are generally going in the right direction.
○ Have gotten off on the wrong track.

What do you think is the most important problem facing this country today?

○ Crime
○ Education
○ The economy
○ Immigration
○ Health care
○ The deficit and government spending
○ Poverty
○ The environment
○ Moral decline
○ The war in Iraq
○ Terrorism
○ The war in Afghanistan
○ Somethings else, please specify

Now thinking about the economy in the country as a whole, would you say that as compared to one year ago, the nation's economy is now better, about the same, or worse?

○



Much better

○ Somewhat better

○ About the same

○ Somewhat worse

○ Much worse

What about 12 months from now? Do you think the economy, in the country as a whole, will be better, about the same, or worse in 12 months?

○ Much better

○ Somewhat better

○ About the same

○ Somewhat worse

○ Much worse

Do you approve of the way each is doing their job...?

	Strongly Approve	Somewhat Approve	Somewhat Disapprove	Strongly Disapprove	Not Sure
President Obama	○	○	○	○	○
The U. S. Congress	○	○	○	○	○
The U. S. Supreme Court	○	○	○	○	○

Do you approve or disapprove of the way Barack Obama is handling each of these issues?

	Strongly Approve	Approve	Disapprove	Strongly Disapprove
Environment	○	○	○	○
Education	○	○	○	○
War on terrorism	○	○	○	○
Economy	○	○	○	○
Taxes	○	○	○	○
Race relations	○	○	○	○
Situation in Iraq	○	○	○	○
Jobs	○	○	○	○
Immigration	○	○	○	○
Energy	○	○	○	○
Health care	○	○	○	○
Foreign policy	○	○	○	○
Situation in Afghanistan	○	○	○	○

When it comes to politics, how would you describe each person or group?

	Very liberal	Somewhat liberal	A little liberal	Neither liberal nor conservative	A little conservative	Somewhat conservative	Very conservative
Yourself	○	○	○	○	○	○	○
Democrats	○	○	○	○	○	○	○
Republicans	○	○	○	○	○	○	○
Barack Obama	○	○	○	○	○	○	○

Sarah Palin	○	○	○	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Congress considered many important bills over the past two years. For each of the following tell us whether you support or oppose the legislation in principle.

	Strongly support	Support	Oppose	Strongly oppose
Financial Reform Bill: Protects consumers against abusive lending. Creates a Bureau of Consumer Financial Protection. Regulates high risk investments known as derivatives. Allow government to shut down failing financial institutions.	○	○	○	○
Federal Intelligence and Security Act: Allow U.S. spy agencies to eavesdrop on overseas terrorist suspects without first getting a court order	○	○	○	○
American Recovery and Reinvestment Act: Authorizes $787 billion in federal spending to stimulate economic growth in the US.	○	○	○	○
Patient Protection and Affordable Care Act: Requires all Americans to have health insurance. Allows people to keep current provider. Sets up health insurance option for those without coverage. Increases taxes on those making more than $280,000 a year.	○	○	○	○
State Children's Health Insurance Program: Program insures children in low income households. Act would renew the program through 2014 and include 4 million additional children.	○	○	○	○
American Clean Energy and Security Act: Imposes a cap on carbon emissions and allow companies to trade allowances for carbon emissions. Funds research on renewable energy.	○	○	○	○
End Don't Ask, Don't Tell: Would allow gays to serve openly in the armed services.	○	○	○	○
Stem Cell Research Enhancement Act: Allow federal funding of embryonic stem cell research.	○	○	○	○

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Now we have a set of questions to help us see how much information about politics gets out to the public. Many people don't know the answers to these questions, but we'd be grateful if you would please answer every question, even if you're not sure what the right answer is.

What job or political office is held by each of the following people?

John Boehner

Joe Biden	
John Roberts, Jr.	
David Cameron	
Jerry Brown	

Which party has a majority of seats in ...

	Republicans	Democrats	Neither	Not Sure
The U. S. House of Representatives	○	○	○	○
The U. S. Senate	○	○	○	○

Whose responsibility is it to determine if a law is constitutional or not?

- ○ The President
- ○ Congress
- ○ The Supreme Court

How much of a majority is required for the U.S. Senate and House to override a presidential veto?

Timing

These page timer metrics will not be displayed to the recipient.
First Click: 0 seconds.
Last Click: 0 seconds.
Page Submit: 0 seconds.
Click Count: 0 clicks.

Demographics

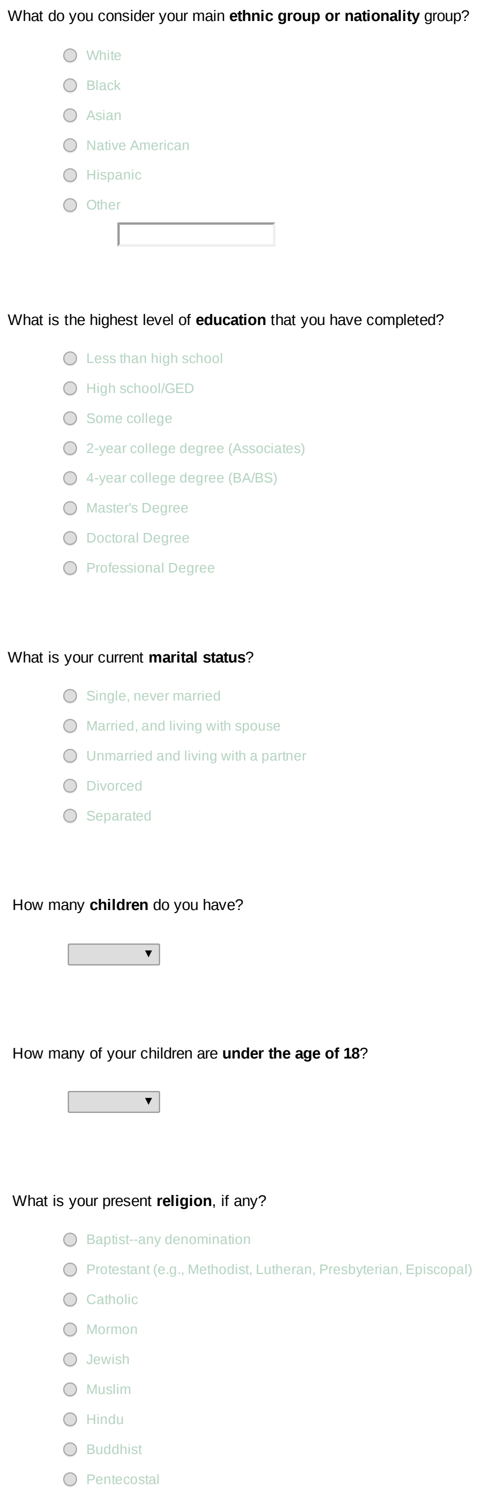
Last, we have some demographic questions.

Are you…?

- ○ Male
- ○ Female

What is your **age**?

▼



What do you consider your main **ethnic group or nationality** group?

- ○ White
- ○ Black
- ○ Asian
- ○ Native American
- ○ Hispanic
- ○ Other

What is the highest level of **education** that you have completed?

- ○ Less than high school
- ○ High school/GED
- ○ Some college
- ○ 2-year college degree (Associates)
- ○ 4-year college degree (BA/BS)
- ○ Master's Degree
- ○ Doctoral Degree
- ○ Professional Degree

What is your current **marital status**?

- ○ Single, never married
- ○ Married, and living with spouse
- ○ Unmarried and living with a partner
- ○ Divorced
- ○ Separated

How many **children** do you have?

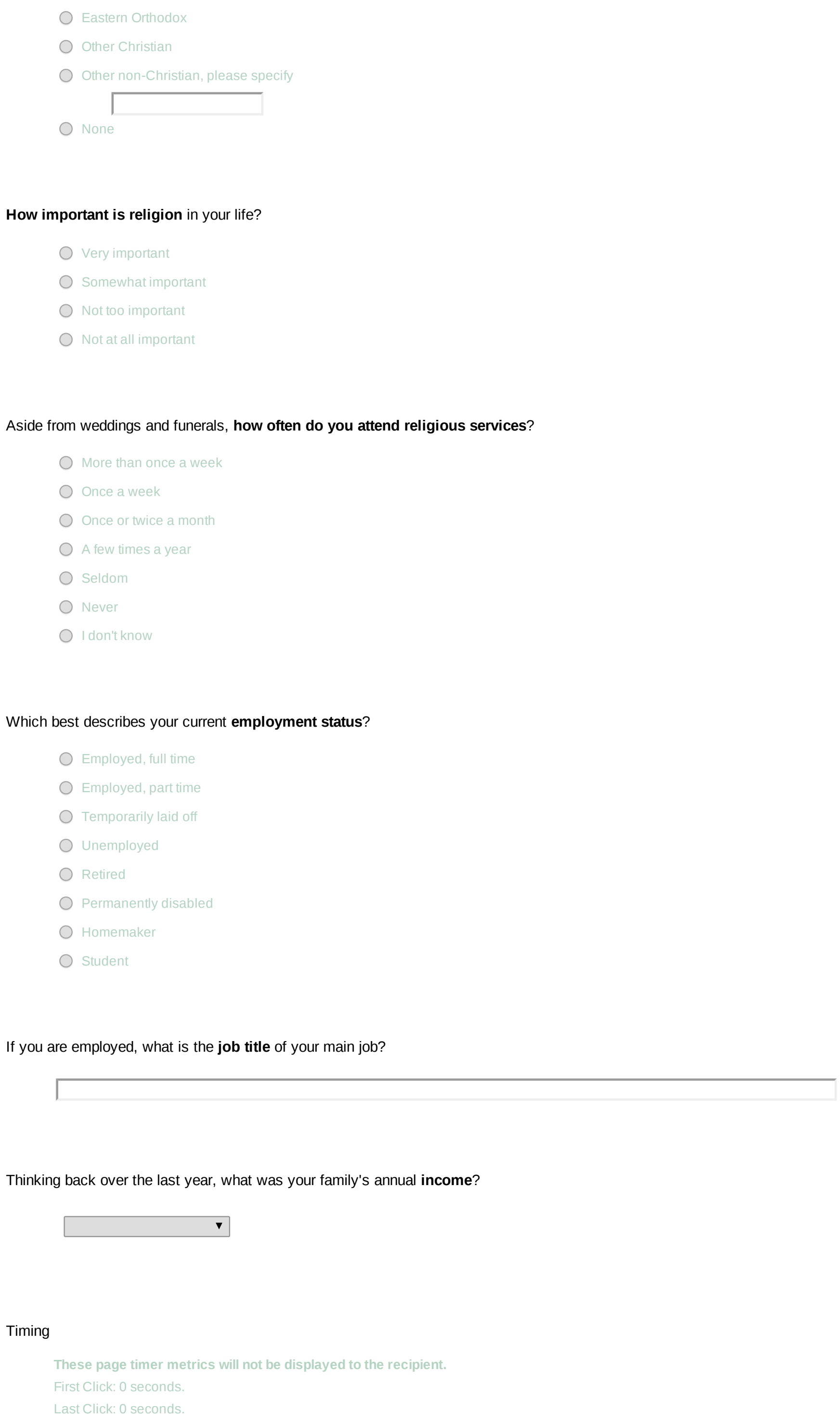
▼

How many of your children are **under the age of 18**?

▼

What is your present **religion**, if any?

- ○ Baptist--any denomination
- ○ Protestant (e.g., Methodist, Lutheran, Presbyterian, Episcopal)
- ○ Catholic
- ○ Mormon
- ○ Jewish
- ○ Muslim
- ○ Hindu
- ○ Buddhist
- ○ Pentecostal



○ Eastern Orthodox

○ Other Christian

○ Other non-Christian, please specify

○ None

How important is religion in your life?

○ Very important

○ Somewhat important

○ Not too important

○ Not at all important

Aside from weddings and funerals, **how often do you attend religious services**?

○ More than once a week

○ Once a week

○ Once or twice a month

○ A few times a year

○ Seldom

○ Never

○ I don't know

Which best describes your current **employment status**?

○ Employed, full time

○ Employed, part time

○ Temporarily laid off

○ Unemployed

○ Retired

○ Permanently disabled

○ Homemaker

○ Student

If you are employed, what is the **job title** of your main job?

Thinking back over the last year, what was your family's annual **income**?

▼

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

Conclusion

Conclusion

Thanks again for your participation. Your responses will help us understand how the blogging community thinks about and responds to important issues and current events. We appreciate your time and opinions.

Please do not blog about this survey -- yet. In order to make sure that the survey experience is consistent for all respondents, we ask you not to blog about or discuss this survey or its contents while it is in the field. Once the survey is complete, we will be happy to share results, and you will be free to talk about it.

□ I would like to be notified once the survey is complete.

Are you willing to participate in similar surveys in the future?

○ Yes, I'm interested in participating in similar surveys in the future.

○ Yes, I'm willing to participate as long as it doesn't take too much time

○ No, I'm not at all interested in similar surveys in the future.

Currently, we have ${e://Field/Email} listed as your email address. If this is not right, please enter your preferred **email address** here.

We welcome **questions and comments**. If you have any questions or comments about the survey, please enter them here.

Thank you!

Timing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds.

Last Click: 0 seconds.

Page Submit: 0 seconds.

Click Count: 0 clicks.

APPENDIX D

Supplementary materials for Mechanical Turk

D.1 FAQ

What is OPSP?

OPSP stands for Online Political Speech Project. We are a University of Michigan research team studying political speech on the Internet. There is no commercial or political agenda behind the project. We're just trying to get a clear picture of how people discuss current events and social issues online.

What bonuses do you award?

We award bonuses based on your accuracy score and the total number of HITs you complete. We try to be generous with bonuses so that quality counts as much as quantity. If your answers are very accurate, you can more than double your pay per HIT.

How much do these HITs pay per hour?

Amazon makes it hard to figure this out in advance, but we've ballparked the hourly rate at \$4 to \$6 base pay, and double that after bonuses. Since bonuses are directly tied to accuracy, the more accurate your answers, the more you get paid.

Where do accuracy scores come from?

We compare your answers to previous answers by experts and other turkers. You get points depending on how closely they match. Since even the experts sometimes disagree, we award lots of partial credit. If your answers are just as good as the experts', your score will be 100 points. If it looks like you're just guessing, your scores will average out close to 0.

How many HITs do I need to complete to get an accuracy score?

20 to 30 HITs will usually be enough. However, the more HITs you do, the more precise your accuracy score will be, so don't hold back!

How often do you calculate accuracy scores?

We compute accuracy scores as fast as we can—usually overnight—so that you don't have to wait too long to find out how you're doing.

Do you ever reject HITs?

Sadly, some turkers try and cheat on HITs by guessing at answers without really reading the articles. We can pick out these cheaters based on low accuracy scores. Our policy is to give a warning, then start rejecting HITs if a turker's accuracy doesn't improve.

The good news is that if you're reading and answering carefully, this should never be a problem.

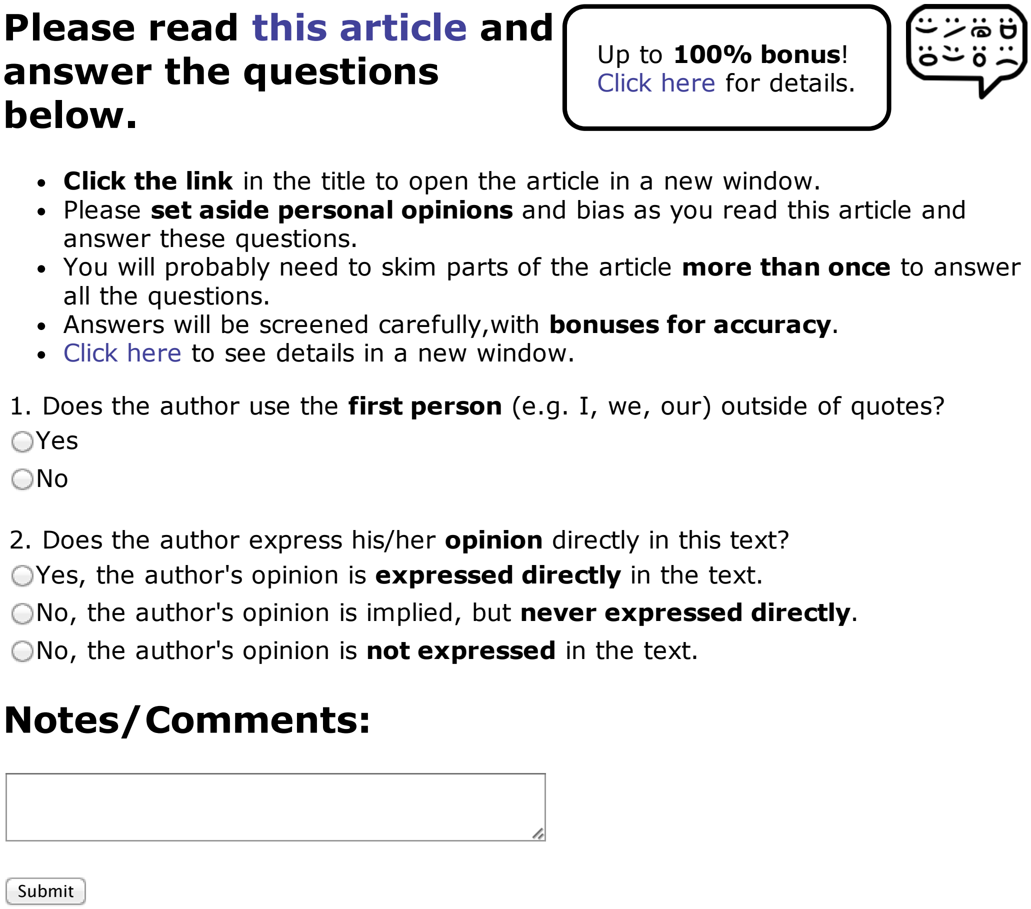
Where do the articles come from?

All over the place. We draw articles from newspapers, transcripts from TV and radio programs, posts from blogs, and comments from discussion boards. All articles are from the public domain.

I like these HITs. Where can I find more?

We're glad you like them! We will post links to the batches here as soon as we can. Also, turkers with high accuracy scores may receive email invitations to new batches as we create them.

D.2 Codebook screen shots



Please read this article and answer the questions below.

Up to **100% bonus**!
Click here for details.

- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

1. Does the author use the **first person** (e.g. I, we, our) outside of quotes?

○Yes

○No

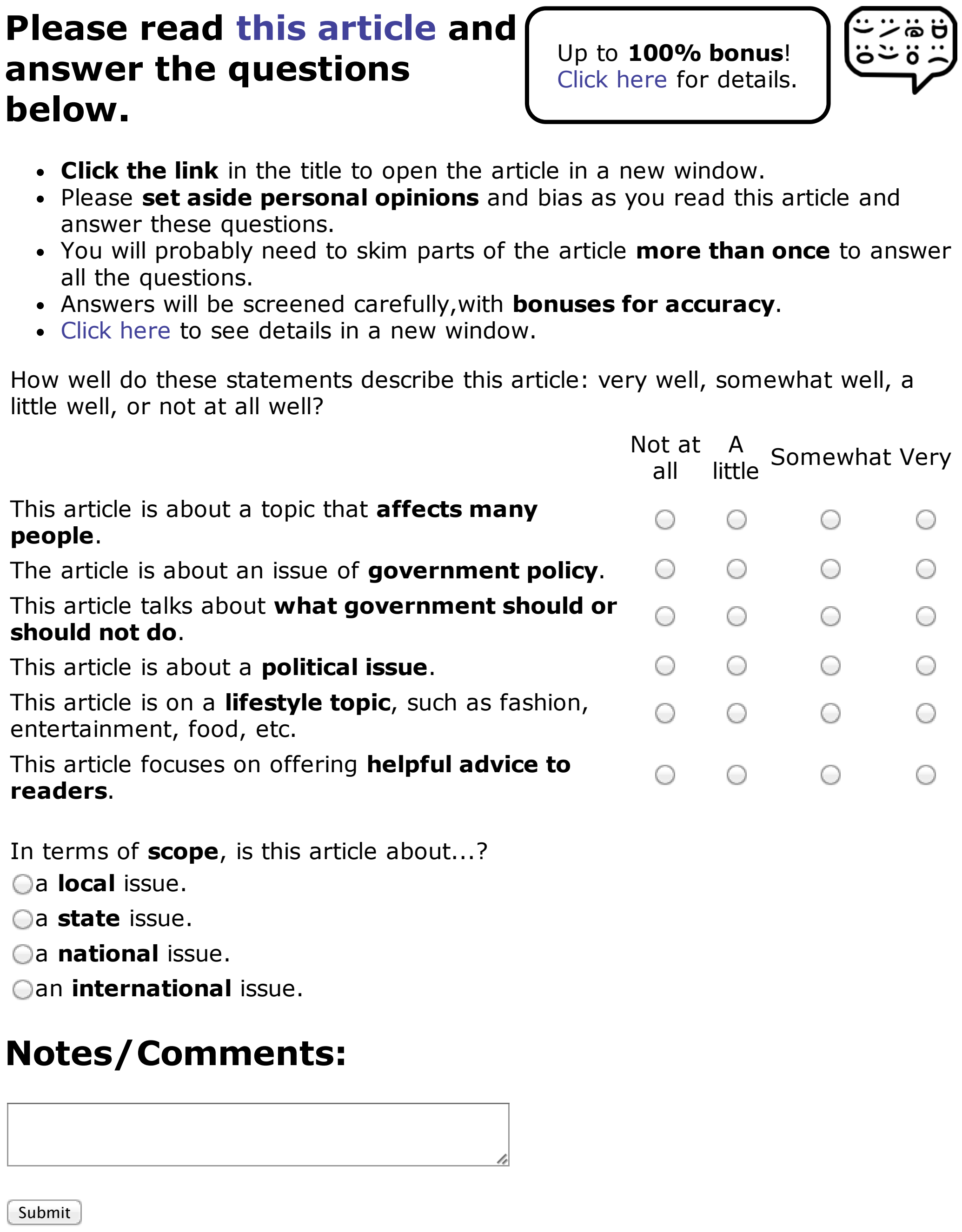
2. Does the author express his/her **opinion** directly in this text?

○Yes, the author's opinion is **expressed directly** in the text.

○No, the author's opinion is implied, but **never expressed directly**.

○No, the author's opinion is **not expressed** in the text.

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**! Click here for details.

- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

How well do these statements describe this article: very well, somewhat well, a little well, or not at all well?

	Not at all	A little	Somewhat	Very
This article is about a topic that **affects many people**.	○	○	○	○
The article is about an issue of **government policy**.	○	○	○	○
This article talks about **what government should or should not do**.	○	○	○	○
This article is about a **political issue**.	○	○	○	○
This article is on a **lifestyle topic**, such as fashion, entertainment, food, etc.	○	○	○	○
This article focuses on offering **helpful advice to readers**.	○	○	○	○

In terms of **scope**, is this article about...?

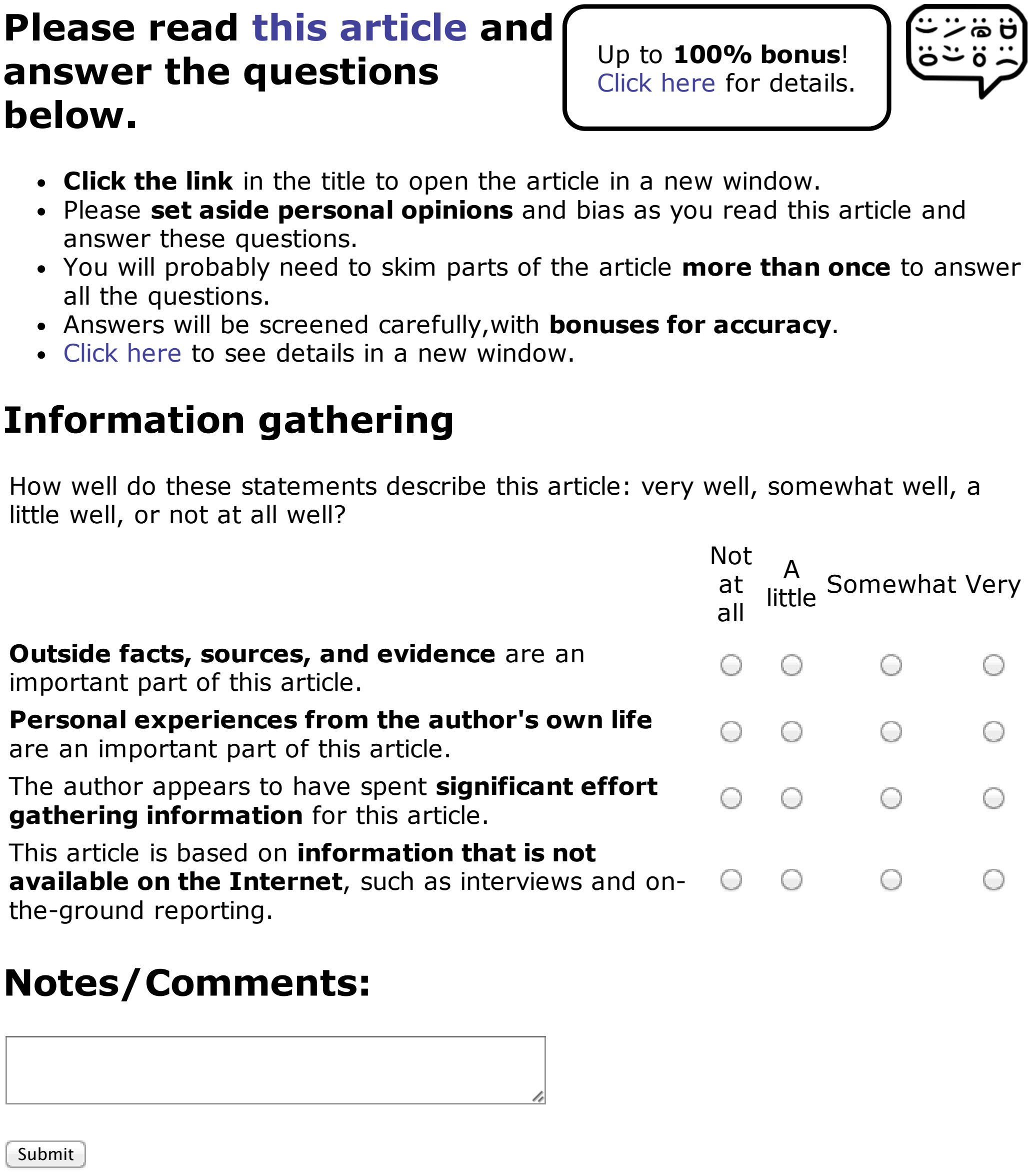
○a **local** issue.

○a **state** issue.

○a **national** issue.

○an **international** issue.

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**! Click here for details.

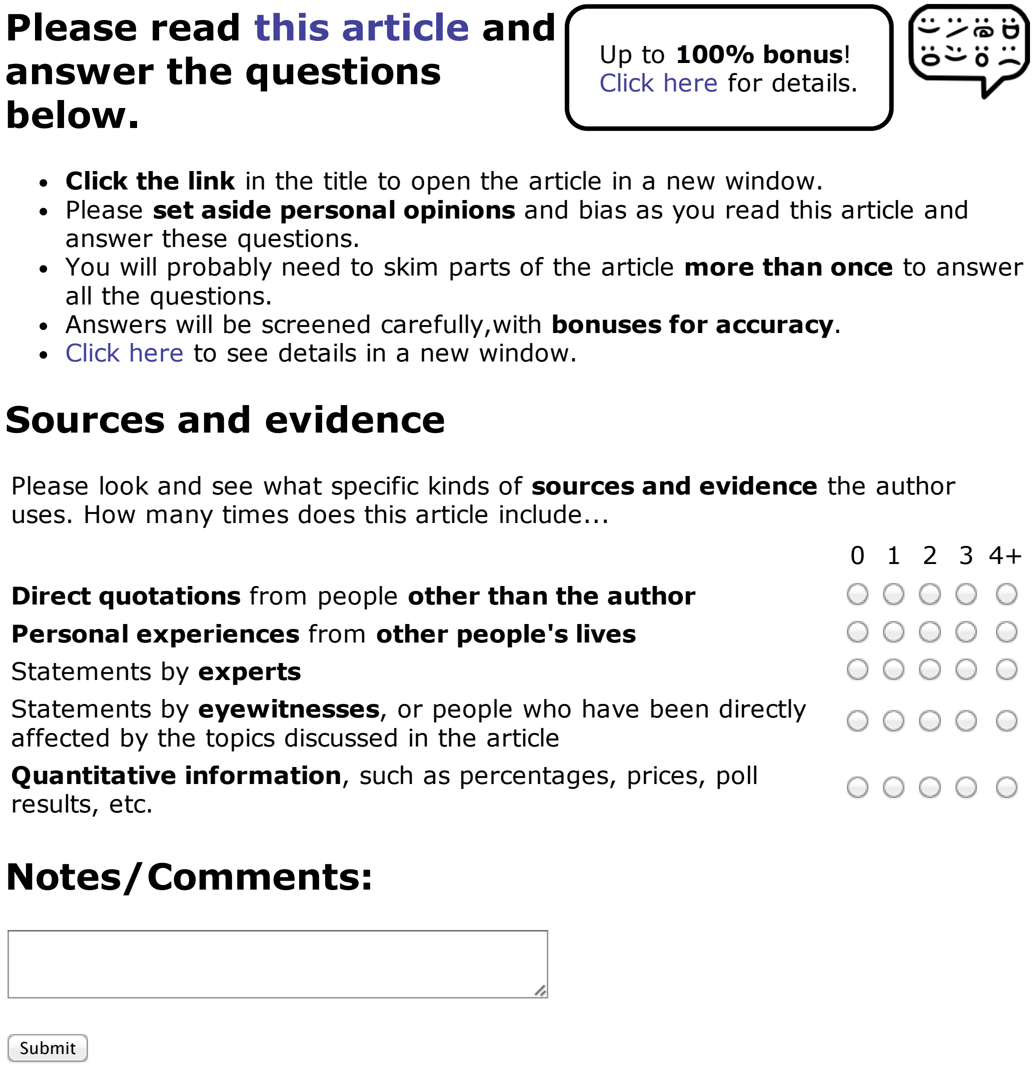
- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

Information gathering

How well do these statements describe this article: very well, somewhat well, a little well, or not at all well?

	Not at all	A little	Somewhat	Very
Outside facts, sources, and evidence are an important part of this article.	○	○	○	○
Personal experiences from the author's own life are an important part of this article.	○	○	○	○
The author appears to have spent **significant effort gathering information** for this article.	○	○	○	○
This article is based on **information that is not available on the Internet**, such as interviews and on-the-ground reporting.	○	○	○	○

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**!
Click here for details.

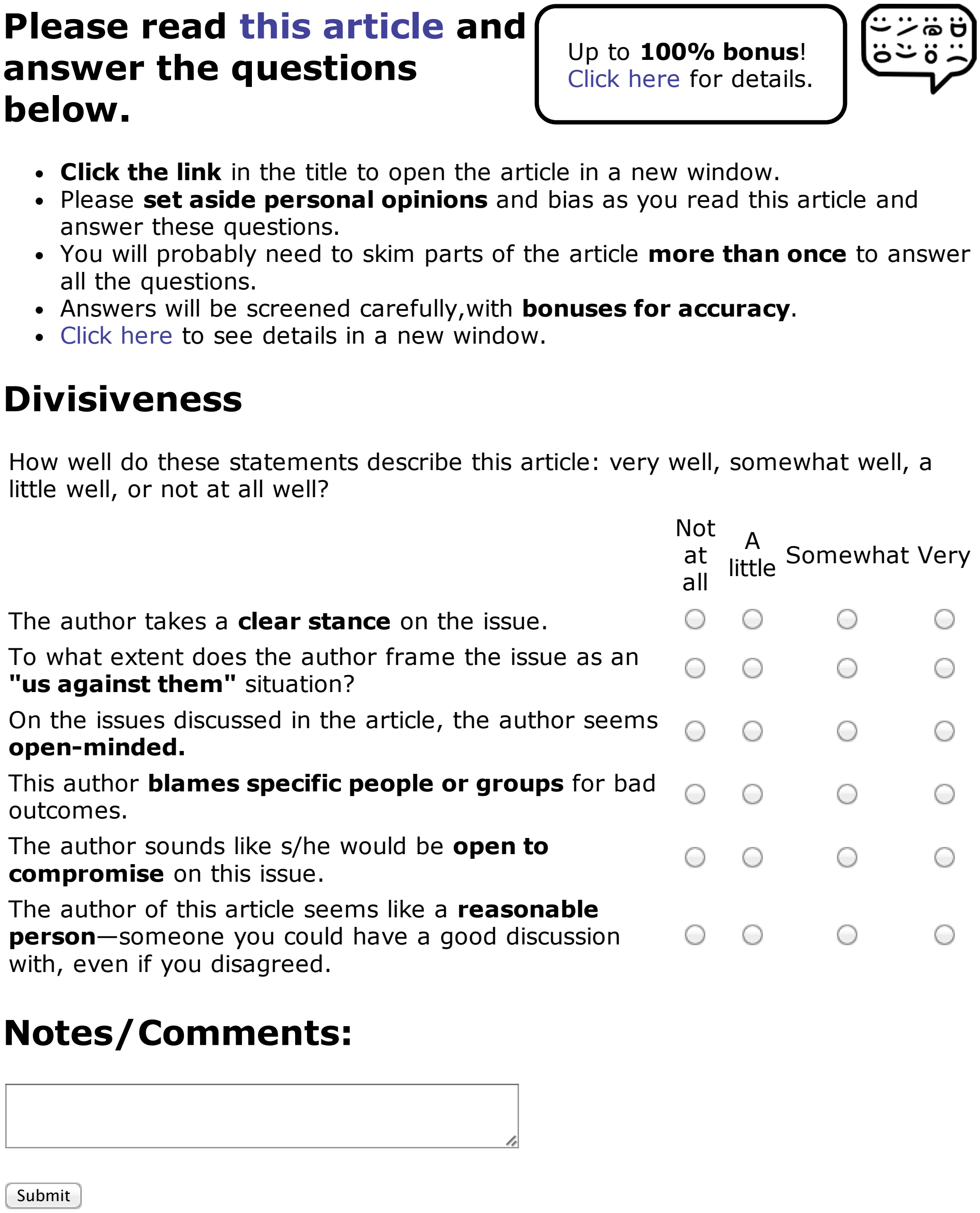
- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

Sources and evidence

Please look and see what specific kinds of **sources and evidence** the author uses. How many times does this article include...

	0	1	2	3	4+
Direct quotations from people **other than the author**	○	○	○	○	○
Personal experiences from **other people's lives**	○	○	○	○	○
Statements by **experts**	○	○	○	○	○
Statements by **eyewitnesses**, or people who have been directly affected by the topics discussed in the article	○	○	○	○	○
Quantitative information, such as percentages, prices, poll results, etc.	○	○	○	○	○

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**!
Click here for details.

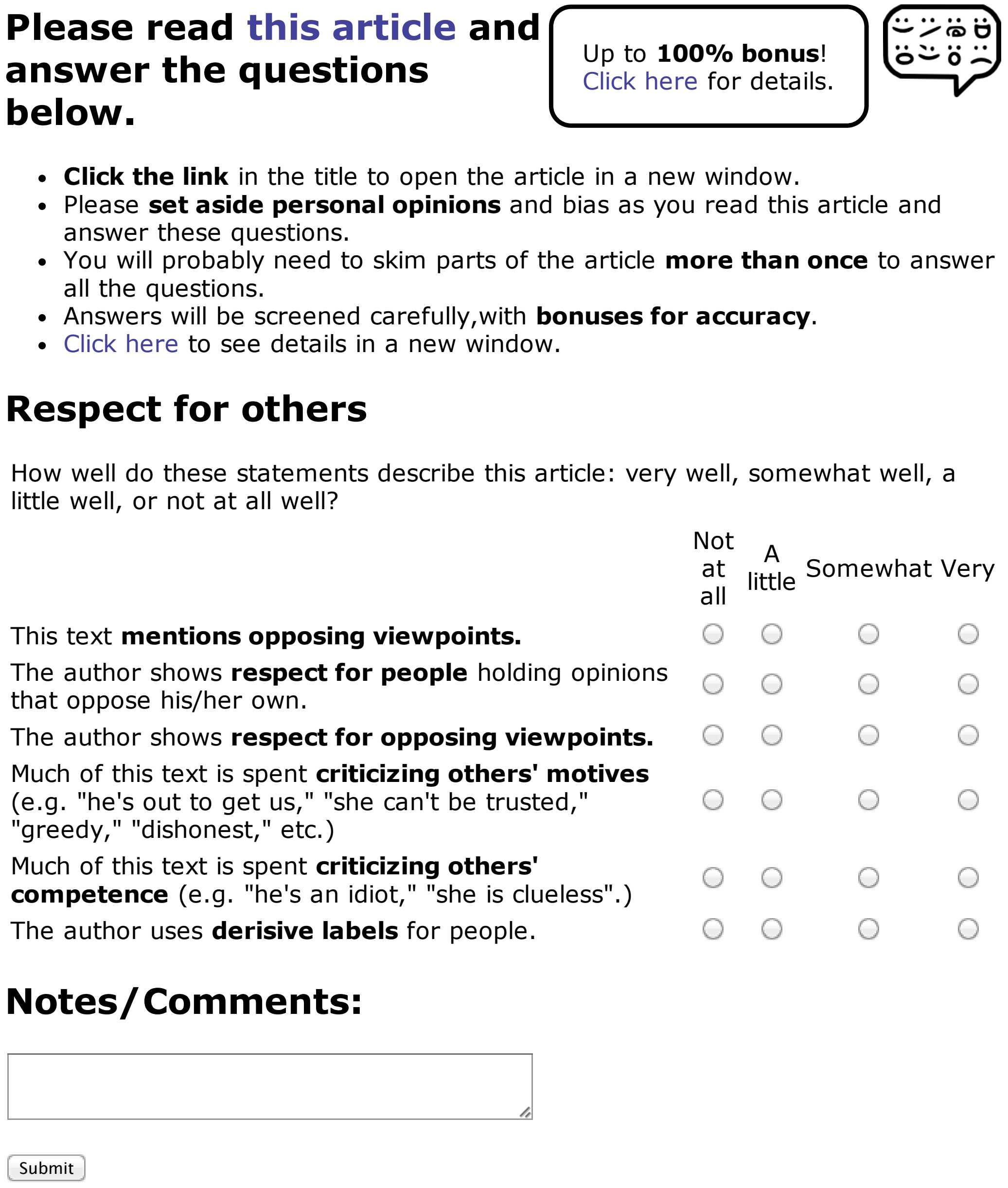
- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

Divisiveness

How well do these statements describe this article: very well, somewhat well, a little well, or not at all well?

	Not at all	A little	Somewhat	Very
The author takes a **clear stance** on the issue.	○	○	○	○
To what extent does the author frame the issue as an **"us against them"** situation?	○	○	○	○
On the issues discussed in the article, the author seems **open-minded.**	○	○	○	○
This author **blames specific people or groups** for bad outcomes.	○	○	○	○
The author sounds like s/he would be **open to compromise** on this issue.	○	○	○	○
The author of this article seems like a **reasonable person**—someone you could have a good discussion with, even if you disagreed.	○	○	○	○

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**!
Click here for details.

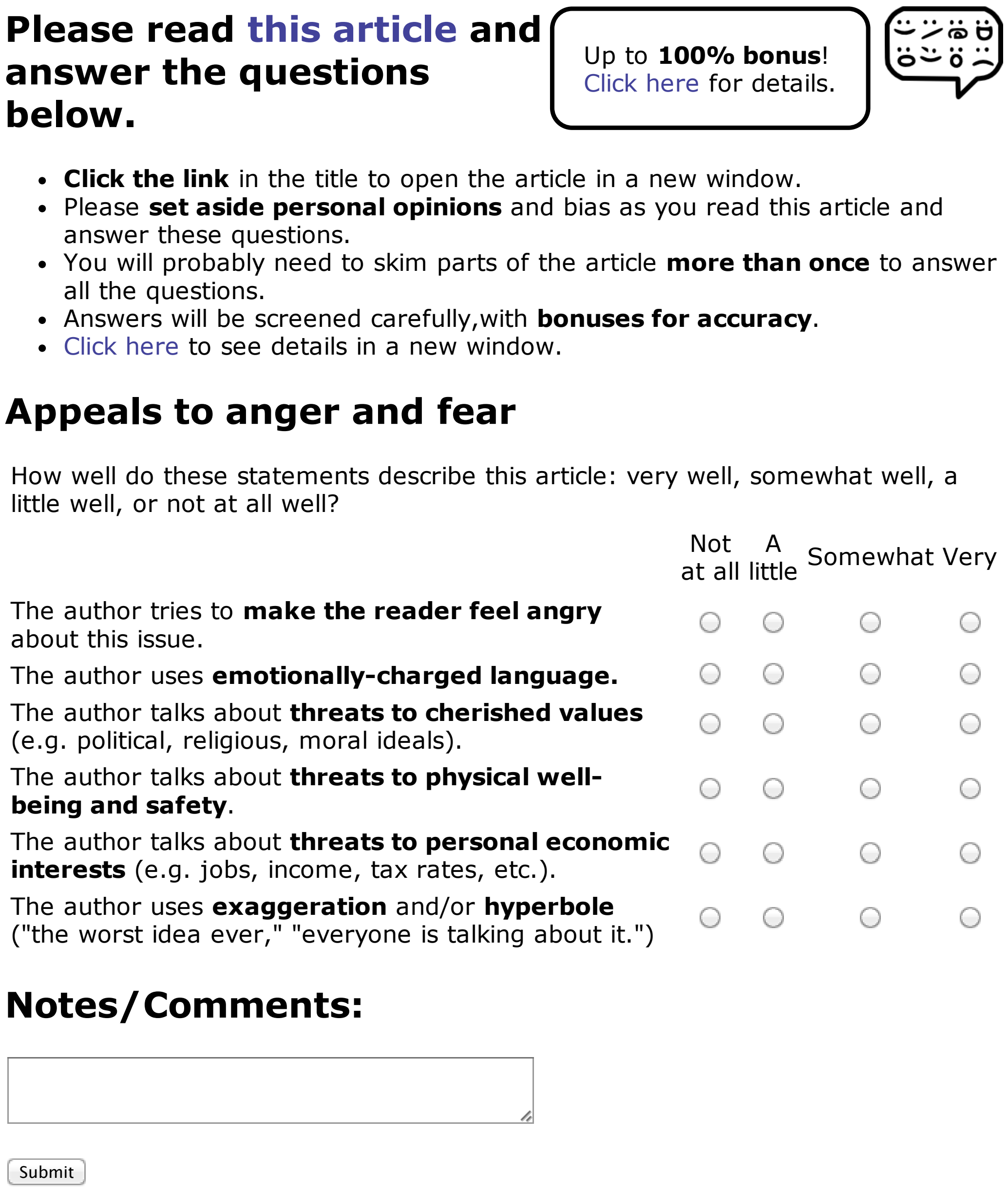
- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

Respect for others

How well do these statements describe this article: very well, somewhat well, a little well, or not at all well?

	Not at all	A little	Somewhat	Very
This text **mentions opposing viewpoints.**	○	○	○	○
The author shows **respect for people** holding opinions that oppose his/her own.	○	○	○	○
The author shows **respect for opposing viewpoints.**	○	○	○	○
Much of this text is spent **criticizing others' motives** (e.g. "he's out to get us," "she can't be trusted," "greedy," "dishonest," etc.)	○	○	○	○
Much of this text is spent **criticizing others' competence** (e.g. "he's an idiot," "she is clueless".)	○	○	○	○
The author uses **derisive labels** for people.	○	○	○	○

Notes/Comments:



Please read this article and answer the questions below.

Up to **100% bonus**!
Click here for details.

- **Click the link** in the title to open the article in a new window.
- Please **set aside personal opinions** and bias as you read this article and answer these questions.
- You will probably need to skim parts of the article **more than once** to answer all the questions.
- Answers will be screened carefully,with **bonuses for accuracy**.
- Click here to see details in a new window.

Appeals to anger and fear

How well do these statements describe this article: very well, somewhat well, a little well, or not at all well?

	Not at all	A little	Somewhat	Very
The author tries to **make the reader feel angry** about this issue.	○	○	○	○
The author uses **emotionally-charged language.**	○	○	○	○
The author talks about **threats to cherished values** (e.g. political, religious, moral ideals).	○	○	○	○
The author talks about **threats to physical well-being and safety**.	○	○	○	○
The author talks about **threats to personal economic interests** (e.g. jobs, income, tax rates, etc.).	○	○	○	○
The author uses **exaggeration** and/or **hyperbole** ("the worst idea ever," "everyone is talking about it.")	○	○	○	○

Notes/Comments:

BIBLIOGRAPHY

BIBLIOGRAPHY

Adamic, L.A. and N. Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery.* ACM pp. 36–43.

Adar, Eytan. 2011. Why I hate Mechanical Turk research (and workshops). In *Proc. CHI Workshop on Crowdsourcing and Human Computation.*

Agresti, A. 2007. *An introduction to categorical data analysis.* Wiley-Blackwell.

Anderson, C. 2008. *Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More.* Hyperion.

Arango, T. 2009. "Fall in newspaper sales accelerates to pass 7%." *The New York Times* 27.

Ashton, Robert H. 1986. "Combining the judgments of experts: How many and which ones?" *Organizational Behavior and Human Decision Processes* 38(3):405–414.

Bauerlein, M., S.G. Walesh et al. 2009. "The Dumbest GenerationHow the Digital Age Stupefies Young Americans and Jeopardizes Our Future." *Leadership and Management in Engineering* 9:100.

Baum, M.A. and T. Groeling. 2008. "New media and the polarization of American political discourse." *Political Communication* 25(4):345–365.

Baumgartner, Frank R, Bryan D Jones, John Wilkerson and E Scott Adler. 2003. "The Policy Agendas Project." *Databases distributed through the University of Washington Center for American Politics. Accessed April* 6:2007.

Baym, Geoffrey. 2005. "The Daily Show: Discursive integration and the reinvention of political journalism." *Political Communication* 22(3):259–276.

Benkler, Y. 2006. *The wealth of networks: How social production transforms markets and freedom.* Yale Univ Pr.

Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2011. "Using Mechanical Turk as a subject recruitment tool for experimental research." *Submitted for review*
.

Berinsky, A.J., G.A. Huber and G.S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Best, S.J. and B.S. Krueger. 2005. "Analyzing the representativeness of Internet political participation." *Political Behavior* 27(2):183–216.

Bimber, Bruce. 2003. *Information and American democracy: Technology in the evolution of political power*. Cambridge University Press.

Bishop, B. 2009. *The big sort: Why the clustering of like-minded American is tearing us apart*. Mariner Books.

BlogPulse.com. 2011. `http://www.blogpulse.com/`.

Blumberg, S.J. 2011. *Wireless Substitution: State-level Estimates from the National Health Interview Survey, January 2007-June 2010*. US Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Bowers, J. 2003. "The dynamics of political participation in the lives of ordinary Americans." *Unpublished doctoral thesis, University of California, Berkeley* .

Buhrmester, Michael, Tracy Kwang and Samuel D Gosling. 2011. "Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1):3–5.

Burns, N., D.R. Kinder and A.M. Ortiz. 2002. Conviction and its Consequences. In *Annual Meeting of the American Political Science Association. Boston, Massachusetts*.

Burns, N., K.L. Schlozman and S. Verba. 2001. *The private roots of public action: Gender, equality, and political participation*. Harvard Univ Pr.

Carpini, M.X.D., F.L. Cook and L.R. Jacobs. 2004. "Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature." *Annu. Rev. Polit. Sci.* 7:315–344.

Carr, N. 2011. *The shallows: What the Internet is doing to our brains*. WW Norton & Company.

Carr, N.G. 2008. *The big switch: Rewiring the world, from Edison to Google*. WW Norton & Company Incorporated.

Chakrabarti, Soumen, Martin Van den Berg and Byron Dom. 1999. "Focused crawling: a new approach to topic-specific Web resource discovery." *Computer Networks* 31(11):1623–1640.

Chen, S. 2009. "Newspapers fold as readers defect and economy sours." *CNN. com* 19.

Clemen, Robert T. 1989. “Combining forecasts: A review and annotated bibliography.” *International Journal of Forecasting* 5(4):559–583.

Cohn, David A, Zoubin Ghahramani and Michael I Jordan. 1996. “Active learning with statistical models.” *arXiv preprint cs/9603104* .

Coleman, S. 2005. “Blogs and the new politics of listening.” *The Political Quarterly* 76(2):272–280.

Conway, Andrew. 2013. “Methods for Collecting Large-Scale Non-Expert Text Coding.” *Available at SSRN* .

Couper, Mick P. 2000. “Review: Web surveys: A review of issues and approaches.” *The Public Opinion Quarterly* 64(4):464–494.

Davis, R. 2009. *Typing politics: the role of blogs in American politics*. Oxford University Press, USA.

Dehejia, Rajeev H and Sadek Wahba. 2002. “Propensity score-matching methods for nonexperimental causal studies.” *Review of Economics and statistics* 84(1):151–161.

Dillman, D.A. 2007. *Mail and internet surveys: The tailored design method*. John Wiley & Sons Inc.

Drezner, D.W. and H. Farrell. 2008. “The power and politics of blogs.” *Public choice* 134(1-2):15–30.

Fiorina, M.P. 1981. “Retrospective voting in American national elections.”.

Fowler, Floyd J. 2009. *Survey research methods*. Vol. 1 Sage.

Franklin, C.H. 1989. “Estimation across data sets: Two-stage auxiliary instrumental variables estimation (2SaIV).” *Political Analysis* 1(1):1.

Frey, D. 1986. “Recent research on selective exposure to information.” *Advances in experimental social psychology* 19:41–80.

Gans, H.J. 1979. “Deciding what’s news: A study of CBS evening news, NBC nightly news, Newsweek, and Time.”.

Gillmor, D. 2006. *We the media: Grassroots journalism by the people, for the people*. O’Reilly Media.

Gong, W. Abraham. 2011. Does the political blogosphere represent or distort the voice of the electorate? In *annual meeting of the Midwest Political Science Association*.

Graf, J. 2006. “The audience for political blogs.” *New research on Blog Readership. GWs Institute for Politics, Democracy & the Internet. Retrieved from: http://www. ipdi. org/uploadedfiles/audience% 20for% 20political% 20blogs. pdf (7/4/2008)* .

Groves, Robert M, Don Dillman, John L Eltinge and Roderick JA Little. 2002. *Survey nonresponse*. Wiley New York.

Gutmann, Amy and Dennis Thompson. 2009. *Why deliberative democracy?* Princeton University Press.

Habermas, J. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. The MIT Press.

Harb, Z. 2011. "Arab Revolutions and the Social Media Effect." *M/C Journal* 14(2).

Hausman, J.A., J. Abrevaya and F.M. Scott-Morton. 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of Econometrics* 87(2):239–269.

Heckathorn, D.D. 1997. "Respondent-driven sampling: a new approach to the study of hidden populations." *Social problems* 44(2):174–199.

Hindman, M. 2010. *The myth of digital democracy*. Princeton University Press.

Hindman, M., K. Tsioutsiouliklis and J.A. Johnson. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *annual meeting of the Midwest Political Science Association*. Vol. 4 Citeseer pp. 1–33.

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

Horton, John J, David G Rand and Richard J Zeckhauser. 2011. "The online laboratory: Conducting experiments in a real labor market." *Experimental Economics* 14(3):399–425.

Hout, Michael. 2012. "Social and economic returns to college education in the United States." *Annual Review of Sociology* 38:379–400.

Howe, J. 2009. *Crowdsourcing*. Random House.
URL: *http://books.google.com/books?id=IMN0dUIxQWIC*

Huckfeldt, Robert and John Sprague. 1992. "Political parties and electoral mobilization: Political structure, social structure, and the party canvass." *The American Political Science Review* pp. 70–86.

Huckfeldt, R.R. and J. Sprague. 1995. *Citizens, politics and social communication: Information and influence in an election campaign*. Cambridge University Press.

Ipeirotis, Panagiotis G, Foster Provost and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM pp. 64–67.

Iyengar, S. and D.R. Kinder. 2010. *News that matters: Television and American opinion*. University of Chicago Press.

Iyengar, Shanto and Kyu S Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of Communication* 59(1):19–39.

Jarvis, J. 2009. *What would Google do?* HarperBusiness.

Jennings, M Kent and Barbara G Farah. 1981. "Social roles and political resources: An over-time study of men and women in party elites." *American Journal of Political Science* pp. 462–482.

Johnson, T.J. and B.K. Kaye. 2004. "Wag the blog: How reliance on traditional media and the Internet influence credibility perceptions of weblogs among blog users." *Journalism and Mass Communication Quarterly* 81:622–642.

Kalmoe, Nathan. 2011. A Call to Arms: How'Fighting'Words Mobilize Political Participation with Aggression. In *APSA 2011 Annual Meeting Paper*.

Karger, David R, Sewoong Oh and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*. pp. 1953–1961.

Karpf, D. 2008. "Understanding blogspace." *Journal of Information Technology & Politics* 5(4):369.

Karpf, D. 2009. The MoveOn Effect: Disruptive Innovation within the Interest Group Ecology of American Politics. In *Paper Presentation at American Political Science Association Annual Meeting, Toronto, CA*.

Katz, Daniel and Samuel J Eldersveld. 1961. "The impact of local party activity upon the electorate." *Public Opinion Quarterly* 25(1):1–24.

Keen, A. 2007. *The cult of the amateur: how today's internet is killing our culture*. Broadway Business.

Key, V.O. 1961. *Public opinion and American democracy*. Knopf New York.

King, G., R.O. Keohane and S. Verba. 1994. *Designing social inquiry*. Princeton University Press Princeton, NJ.

Kittur, Aniket, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM pp. 1301–1318.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Sage Publications, Inc.

Lamberson, PJ and Scott E Page. 2012. "Optimal forecasting groups." *Management Science* 58(4):805–810.

Lasica, J.D. 2003. "Blogs and journalism need each other." *Nieman Reports* 57(3):70–74.

Laver, M., K. Benoit and J. Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Lawrence, E., J. Sides and H. Farrell. 2010. "Self-segregation or deliberation? Blog readership, participation, and polarization in American politics." *Perspectives on Politics* 8(01):141–157.

Lazear, Edward P. 1995. *Personnel economics (series).* Vol. 1993 the MIT Press.

Lazer, D., A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann et al. 2009. "Life in the network: the coming age of computational social science." *Science (New York, NY)* 323(5915):721.

Lee, Sunghee. 2006. "Propensity score adjustment as a weighting scheme for volunteer panel web surveys." *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-* 22(2):329.

Lenhart, A. and S. Fox. 2006. "Pew Internet Report–Bloggers: A portrait of the internet\'s new storytellers.".

Leskovec, J., L. Backstrom and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* Citeseer pp. 497–506.

Li, Dan. 2005. *Why do you blog: A uses-and-gratifications inquiry into bloggers' motivations.* Vol. 17 Citeseer.

Lippmann, W. 1927. *Public opinion.* Macmillan.

Liu, Qiang, Jian Peng and Alex Ihler. 2012. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems.* pp. 701–709.

MacKuen, M.B., R.S. Erikson and J.A. Stimson. 1989. "Macropartisanship." *The American Political Science Review* pp. 1125–1142.

Manning, C.D., P. Raghavan, H. Schutze and Ebooks Corporation. 2008. *Introduction to information retrieval.* Vol. 1 Cambridge University Press Cambridge, UK.

Maron, ME. 1961. "Automatic indexing: an experimental inquiry." *Journal of the ACM (JACM)* 8(3):404–417.

Mason, Winter and Duncan J Watts. 2010. "Financial incentives and the performance of crowds." *ACM SigKDD Explorations Newsletter* 11(2):100–108.

Mayhew, D.R. 2004. *Congress: The electoral connection.* Vol. 26 Yale Univ Pr.

McKenna, L. and A. Pole. 2008. “What do bloggers do: an average day on an average political blog.” *Public Choice* 134(1):97–108.

McNamee, J Paul, James C Mayfield, Martin R Hall, Lien T Duong and Christine D Piatko. 2002. “Directed web crawler with machine learning.”. US Patent App. 10/121,525.

Menczer, Filippo, Gautam Pant and Padmini Srinivasan. 2004. “Topical web crawlers: Evaluating adaptive algorithms.” *ACM Transactions on Internet Technology (TOIT)* 4(4):378–419.

Merton, Robert K. 1968. “Social theory and social structure.”.

Meyer, P. 2009. *The vanishing newspaper: Saving journalism in the information age.* University of Missouri.

Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2008. Coder reliability and misclassification in comparative manifesto project codings. In *66th MPSA Annual National Conference.* pp. 3–6.

Miller, E.A., A. Pole and C. Bateman. 2011. “Variation in Health Blog Features and Elements by Gender, Occupation, and Perspective.” *Journal of health communication* 99999(1):1–24.

Morozov, Evgeny. 2012. *The net delusion: The dark side of Internet freedom.* PublicAffairs Store.

Mutz, D.C. 2006. *Hearing the other side: Deliberative versus participatory democracy.* Cambridge University Press.

Mutz, Diana C. 2008. “Is deliberative democracy a falsifiable theory?” *Annu. Rev. Polit. Sci.* 11:521–538.

Negroponte, N. 1995. *Being digital.* Alfred A. Knopf.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook.* Sage.

Ng, A.Y. 2004. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning.* ACM p. 78.

Nie, Norman H. 1996. *Education and democratic citizenship in America.* University of Chicago Press.

Noren, Laura. 2012. “Food blog study.”.
URL: *http://www.foodblogstudy.info/*

Ogilvy, Public Relations and University Center for Social Impact Communication Georgetown. 2011. “Dynamics of cause engagement.”.

Page, Benjamin I and Robert Y Shapiro. 2010. *The rational public: Fifty years of trends in Americans' policy preferences*. University of Chicago Press.

Perlmutter, D.D. 2008. "Blogwars: The new political battleground.".

Pole, A. 2009. *Blogging the political: politics and participation in a networked society*. Taylor & Francis.

Porter, M. 2006. "Porter Stemming Algorithm, 2006." *URL: http://www. tartarus. org/martin/PorterStemmer, Accessed on March* 20.

Postman, N. 1993. *Technopoly: The surrender of culture to technology*. Vintage.

Prelec, Dražen. 2004. "A Bayesian truth serum for subjective data." *Science* 306(5695):462–466.

Prior, M. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge Univ Pr.

Reno, R.R., R.B. Cialdini and C.A. Kallgren. 1993. "The transsituational influence of social norms." *Journal of personality and social psychology* 64(1):104.

Reynolds, G. 2007. *An army of Davids: How markets and technology empower ordinary people to beat big media, big government, and other Goliaths*. Thomas Nelson Inc.

Rheingold, H. 2003. *Smart mobs: The next social revolution*. Basic books.

Rivers, D. 2007. Sampling for web surveys. In *Joint Statistical Meetings*.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Rosenstone, S. and J.M. Hansen. 1993. "Mobilization, participation and democracy in America.".

royal.pingdom.com. 2011. "Internet 2010 in numbers.".
URL: *http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/*

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological methodology* 34(1):193–240.

Sanders, Lynn M. 1997. "Against deliberation." *Political theory* 25(3):347–376.

Scherer, Michael. 2012. "Exclusive: Obamas 2012 Digital Fundraising Outperformed 2008.".
URL: *http://swampland.time.com/2012/11/15/exclusive-obamas-2012-digital-fundraising-outperformed-2008/*

Schlesselman, J.J. and P.D. Stolley. 1982. *Case-control studies: design, conduct, analysis*. Vol. 2 Oxford University Press, USA.

Schlozman, Kay Lehman, Nancy Burns and Sidney Verba. 1999. "What happened at work today?: A multistage model of gender, employment, and political participation." *The Journal of Politics* 61(01):29–53.

Schlozman, K.L., S. Verba and H.E. Brady. 2010. "Weapon of the strong? Participatory inequality and the Internet." *Perspectives on Politics* 8(2):487–509.

Schmitt, J.C. 1991. "Trigram-based method of language identification.". US Patent 5,062,143.

Schudson, M. 1981. *Discovering the news: A social history of American newspapers.* Basic Books.

Schudson, M. 2001. "The objectivity norm in American journalism*." *Journalism* 2(2):149–170.

Sears, D.O. and J.L. Freedman. 1967. "Selective exposure to information: A critical review." *Public Opinion Quarterly* 31(2):194.

Shannon, C.E. and W. Weaver. 1949. "The mathematical theory of information.".

Shaw, Aaron D, John J Horton and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work.* ACM pp. 275–284.

Sheng, Victor S, Foster Provost and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM pp. 614–622.

Shirky, C. 2009. *Here comes everybody.* Penguin Books.

Smith, A. 2008. "New numbers for blogging and blog readership." *Pew Internet Posts* .

Sobieraj, Sarah and Jeffrey M Berry. 2011. "From incivility to outrage: Political discourse in blogs, talk radio, and cable news." *Political Communication* 28(1):19–41.

Soon, C. and H. Cho. 2011. "Flows of Relations and Communication among Singapore Political Bloggers and Organizations: The Networked Public Sphere Approach." *Journal of Information Technology & Politics* 8(1):93–109.

Stimson, J.A. 2004. *Tides of consent: How public opinion shapes American politics.* Cambridge Univ Pr.

Sunstein, C.R. 2007. *Republic. com 2.0.* Princeton Univ Pr.

Surowiecki, J. 2005. *The wisdom of crowds.* Anchor.

Tapscott, D. and A.D. Williams. 2008. *Wikinomics: How mass collaboration changes everything.* Portfolio Trade.

Tilly, C. 1978. *From mobilization to revolution.* McGraw-Hill New York.

Tilly, C. 2004. *Social Movements, 1768-2004.* Paradigm Publishers Boulder, CO.

Verba, S., K.L. Schlozman and H.E. Brady. 1995. *Voice and equality: Civic voluntarism in American politics.* Cambridge Univ Press.

Verba, S. and N.H. Nie. 1987. *Participation in America: Political democracy and social equality.* University of Chicago Press.

Wallsten, K. 2008. "Political blogs: Transmission belts, soapboxes, mobilizers, or conversation starters?" *Journal of Information Technology & Politics* 4(3):19–40.

Walsh, K.C. 2004. *Talking about politics: Informal groups and social identity in American life.* University of Chicago Press.

Weisberg, Herbert F. 2009. *The total survey error approach: A guide to the new science of survey research.* University of Chicago Press.

Welinder, Peter, Steve Branson, Pietro Perona and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems.* pp. 2424–2432.

White, D. and P. Winn. 2009. "State of the blogosphere 2008." *Technorati Report* .

Wolfinger, R.E. and S.J. Rosenstone. 1980. *Who votes?* Vol. 22 Yale Univ Pr.

Woodly, D. 2008. "New competencies in democratic communication? Blogs, agenda setting and political participation." *Public Choice* 134(1):109–123.

Zaller, J. 1999. "A theory of media politics." *Manuscript, October* 24:1999.

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

Zhang, T. and F.J. Oles. 2001. "Text categorization based on regularized linear classification methods." *Information Retrieval* 4(1):5–31.

Zhou, Daniel Xiaodan. 2013. "Liberal or conservative: Evaluation and classification with distribution as ground truth.". Dissertation manuscript.

Zhou, Dengyong, John Platt, Sumit Basu and Yi Mao. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25.* pp. 2204–2212.

Zittrain, J. 2009. *The future of the internet–and how to stop it.* Yale University Press.