Open Content, Linus' Law, and Neutral Point of View

Citation

Published Version

Permanent link

Terms of Use

Share Your Story

Open Content, Linus' Law, and Neutral Point of View*

Shane Greenstein
Harvard Business School
Boston, MA 02163
sg@hbs.edu

Feng Zhu
Harvard Business School
Boston, MA 02163
fzhu@hbs.edu

March 5, 2016

Open Content, Linus' Law, and Neutral Point of View

Abstract

The diffusion of the Internet and digital technologies has enabled many organizations to use the open-content production model to produce and disseminate knowledge. While several prior studies have shown that the open-content production model can lead to high-quality output in the context of uncontroversial and verifiable information, it is unclear whether this production model will produce any desirable outcome when information is controversial, subjective, and unverifiable. We examine whether the open-content production model helps achieve a neutral point of view (NPOV) using data from Wikipedia's articles on US politics. Our null hypothesis builds on Linus' Law, often expressed as "Given enough eyeballs, all bugs are shallow." Our findings are consistent with a narrow interpretation of Linus' Law, namely, a greater number of contributors to an article makes an article more neutral. No evidence supports a broad interpretation of Linus' Law. Moreover, several empirical facts suggest the law does not shape many articles. The majority of articles receive little attention, and most articles change only mildly from their initial slant. Our study provides the first empirical evidence on the limit of Linus' Law. While many organizations believe that they could improve their knowledge production by leveraging communities, we show that in the case of Wikipedia, there are aspects, such as NPOV, that the community does not always achieve successfully.

1. Introduction

The diffusion of the Internet and digital technologies has significantly increased the number of opportunities for individuals around the globe to collaborate with each other. As a result, an increasing amount of content today is created by online communities, and is updated and disseminated in real time. For example, Wikis, defined as "sets of dynamically created web pages with content contributed directly by users in a web browser" (Yates et al. 2010), are routinely used to capture and share knowledge in the general public. They are also frequently used by corporations for knowledge development and management (Wagner and Majchrzak 2007). In a similar vein, OpenIDEO, an online innovation platform, leverages communities to create solutions for social issues (Lakhani et al. 2012).

The rise of open content in the digital economy has generated tremendous interest in academic literature. Prior studies have examined the social networks behind patterns of contributions in open content projects (Piskorski and Gorbetai 2010; Zhang and Zhu 2011; Ransbotham and Kane 2011; Kane and Ransbotham 2013a), the dynamics of contributions (Chi et al. 2007; Halfaker et al. 2009; Kane and Ransbotham 2013b), the allocation of effort among topics (Gorbetai 2011), and the impact of open content on the market information environment (Xu and Zhang 2013).[1] Studies have also examined the accuracy of content produced from the open content product model. Giles (2005) and Rajagopalan et al. (2011), for example, find that Wikipedia, the largest encyclopedia on the Web, is about as good a source of accurate scientific information as Britannica, an encyclopedia authored by experts.

While these prior studies find that open content performs well in the context of uncontroversial and verifiable information, it is unclear whether an open-content production model will produce any desirable outcome when information is controversial, subjective, and unverifiable. In this study, we examine whether the open content model achieves a neutral point of view (NPOV) for knowledge with such properties. This issue is particularly important in light of the increasing usage of open content as knowledge sources in the educational settings. For example, Wikipedia has been used extensively by students because it offers a

---

[1] See a list of academic studies about Wikipedia at http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia (accessed June 2015).

mixture of coverage, convenience, and comprehensibility.[2] A number of efforts, such as the Open Book Project and Wikibooks[3], seek to leverage communities to develop freely distributable textbooks and educational materials on a wide range of topics. Studies of news media have shown that media bias may significantly influence audience behavior even when a large fraction of the audience is aware of the existence of media bias (e.g., DellaVigna and Kaplan 2007). One would thus expect biases contained in open content to play a more significant role in shaping beliefs and ideologies when such open content is used as educational sources.

We study whether the open content model achieves NPOV in the context of articles about US politics on Wikipedia. NPOV is one of the tenets that all Wikipedia articles aspire to achieve, along with "verifiability" and "the absence of original research." If an article reflects NPOV, then conflicting opinions are presented next to one another, with all significant points of view represented. One would expect that NPOV should not be difficult to achieve when articles cover uncontroversial topics loaded with objective information that can be verified against many sources. Such a setting characterizes the vast majority of Wikipedia articles about established scientific topics, for example. What about topics lacking these ideal features? What biases arise in topics where information is controversial, subjective, and unverifiable? In the context of Wikipedia, although most contributors try to diffuse issues with a fair representation, collective intelligence bias (CIB), the opposite of NPOV, may arise for a number of reasons. For example, some issues are simply too complex for contributors to resolve, such as in the case of interpreting the science behind global warming. Anyone can verify the same objective data, but generating a consensus takes considerable effort and expertise. CIB can also survive because of the difficulties involved in editing subjective information that is costly to verify. Finally, CIB may arise when ideologues try to influence readers' viewpoints by adding biased content.

---

[2]   Sources: http://www.theguardian.com/education/2013/may/13/should-university-students-use-wikipedia and http://firstmonday.org/article/view/2830/2476, accessed August 2015.

[3] Sources: http://openbookproject.net/ and https://www.wikibooks.org, accessed August 2015.

The study sheds light on Wikipedia's revision processes. Our null hypothesis builds on Linus' Law, often expressed as "Given enough eyeballs, all bugs are shallow" (Raymond 1998). According to a narrow interpretation of Linus' Law, articles should come closer to NPOV as more contributors scrutinize them and make contributions. In a broad interpretation, a more widely dispersed set of contributors also should contribute to the production of NPOV.

The near-decade of history at Wikipedia provides sufficient variance to test Linus' Law. With Wikipedia's size and scale, not all articles receive the same amount of attention and contributions. Some have accumulated many contributions over time, while others have not. Articles also vary in the concentration of contributors they attract. In addition, Wikipedia retains prodigious records of its revisions, which allow for a detailed statistical analysis.

We apply the null hypothesis to entries on Wikipedia about US political topics. We select these articles for two reasons. First, achieving NPOV can be challenging when articles cover controversial topics, and rely on subjective information that is costly to verify. Hence, we presume that Linus' Law has its highest probability of failing for these articles. Alternatively, if Linus' Law succeeds here, it is likely to succeed for other topics. Second, it is feasible to measure slant and bias, building on an approach pioneered by Gentzkow and Shapiro (2010), and the broader literature examining content bias. We will define the measures precisely below. Briefly, slant is a number on a negative/positive yardstick, and bias is the absolute value of that number. The gains from having a defined yardstick come with one drawback. Our bias measure is not available for all articles about US politics, so we must consider whether zero slant and bias signals merely lack of information or sample selection issues.

Wikipedia's history also provides some interesting context for this study. Greenstein and Zhu (2012) show that in its earliest years, Wikipedia's political entries lean Democrat, on average, and tend to be biased. Both of these traits diminish over time. By the most recent observation, on average, Wikipedia's entries lack much slant and contain (moderately) less bias than observed earlier. What role did revision play in these trends? Oversimplifying somewhat, if Linus' Law holds, then older articles could lose their slant through more revision, diminishing bias and slant. Alternatively, if Linus' Law does not hold, then another

mechanism, such as the entry of articles with opposite slant, is responsible for the aggregate decline in average slant over time.

This study shows that the evidence supports a narrow interpretation of Linus' Law. Only one feature of the revision process—namely, number of reviewers—shapes the slant or bias of an article in the direction that the Law predicts. Moreover, several facets of the revision process do not shape revisions in the anticipated direction, and no evidence supports a broad interpretation. The evidence further points to the persistence of bias in many articles, potentially consistent with the presence of CIB. This is partly a vintage effect and partly a result of the topic covered by the article. Some topics, such as entries on civil rights, tend to lean Democrat, and some, such as trade, lean Republican. Finally, we show that a fundamental resource issue constrains the implementation of Linus' Law, namely, due to not enough contributors, only a small percentage of articles ever get enough contributions to enable the law to have an opportunity to work.

Readers who are interested in open source or open content product will take an interest in this study. The vast majority of research on open source or open content examines motivations of contributors (e.g., Hann et al. 2004; Singh et al. 2011; Susarla et al. 2012; Hann et al. 2013), the process of production (e.g., MacCormack et al. 2006; Haefliger et al. 2008; Ghose et al. 2012), or determinants of project successes (e.g., Chen et al. 2013; Wen et al. forthcoming), and few examine Linus' Law directly. Also, little work considers content production from an aggregation of contributions from a large numbers of volunteers, as observed in Wikipedia. Wikipedia is a natural subject for the open source community to examine because of the attention it receives. In most countries with developed Internet sectors, Wikipedia ranks among the top-ten websites visited by households.[4] In the US, Wikipedia is one of the most popular web sites in which user-generated content plays a prominent role.

Our research also contributes to the broad literature examining content bias. Scholars have identified various sources of bias in media content, such as pressure from advertisers or the government (e.g., Price 2003; Besley and Prat 2006; Reuter and Zitzewitz 2006), the media's partisan bias (Larcinese et al. 2007),

---

[4] See the rankings at Alexa.com. Wikipedia is the fifth or sixth most-visited web site in the United States, behind Google, Facebook, Yahoo, YouTube, and, perhaps, eBay (accessed May 2015).

and readers' desire to reinforce their prior beliefs (e.g., Groseclose and Milyo 2005; Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2006; Bernhardt et al. 2008; Gal-Or et al. 2010; Gentzkow and Shapiro 2010). Unlike these prior studies, we examine user-generated content and focus on factors that cause content bias to change over time.

We also provide empirical evidence on whether the Internet is increasing ideological segregation (e.g., Sunstein 2001; Carr 2008; Lawrence et al. 2010; Gentzkow and Shapiro 2011). Our results support the view that prominent articles are not isolated—for these articles, contributors with different political viewpoints have dialogues with each other, and that diminishes the slant of articles. In addition, the general movement in Wikipedia's overall slant suggests entry of new opinions is not precluded. On the other hand, most Wikipedia articles only mildly change their slant, consistent with the view they might receive more attention from readers with similar viewpoints.

Our study is the first to develop statistical tests for whether the open content production model achieves NPOV, and to translate widely discussed ideas about Linus' Law into testable propositions. It is also the first to raise questions about limitations of Linus' Law, and, related, this study frames several new open questions about revision, such as feedbacks between an article's bias and further contributions.

The rest of the paper is organized as follows. We first provide a background of Wikipedia in Section 2 and discuss the role of NPOV on Wikipedia in Section 3. We then develop our hypotheses in Section 4. We present our data and summary statistics in Section 5. Section 6 presents our regression results, and Section 7 concludes.

2. The Emergence of Wikipedia

The first wiki was developed in 1995 by Ward Cunningham, a software engineer from Portland, Oregon. Wikis were first developed and intended for documenting software development. Says Cunningham (Levine, 2006), "It's a medium that allows people to collaborate more easily than they could in systems that are modeled after the pre-computer world, like e-mail."

This study examines Wikipedia just prior to its tenth birthday. Wikipedia was founded in 2001 when Wikipedia began to position itself as "the free encyclopedia that anyone can edit," that is, as an online encyclopedia that is entirely written and edited through user contributions (e.g., Greenstein and Devereux 2009). As of September 2015, it is the largest Wiki on the planet: It supports 5.0 million articles in English and well over 37 million articles in all languages, and hosts content that hundreds of millions of readers view each month.

Wikipedia's production defies simple characterization. Because it relies so heavily on user-generated content, it does not fit existing models of production, in which a fixed sequence of activities produces an output following a pre-specified design. Instead, Wikipedia uses a commons-based approach to aggregate and revise information from a widely dispersed set of contributors and it produces non-proprietary information.

Since 2003, Wikipedia has been owned and administered by the Wikimedia Foundation, a not-for-profit group established to manage the operations behind the Wikipedia Web site and related efforts. Until 2006 the foundation operated with a minimal staff of two programmers, under the supervision of Jimmy Wales, but by 2010, the staff had grown to include a full-time professional manager and several dozen employees. Virtually all the content continues to come from volunteers.

Wikipedia operates under an open-source license. When Wikipedia first began, most images and other content were covered by the GNU free documentation license (GFDL), a variant on the more popular GPL designed for manuals, textbooks, and reference materials. With the latter, contributions remained the property of their creators, whereas the GFDL license ensured the content would remain freely distributable and reproducible. More recently, most content is dual-licensed under both the GFDL and/or the Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA).[5] Copies can be sold commercially, but if produced in larger quantities, then the original document or source code must be made available.

---

[5] http://en.wikipedia.org/wiki/Wikipedia:About (accessed August 2015).

Wiki server technology allows the creation of hypertexts with nonlinear navigation structures. Each page contains a series of cross-links to other pages. The reader decides how to navigate through the site. Contributing to Wikipedia is easy and transparent. Contributors do not need specialized knowledge.

As there is no editorial control from the center, Wikipedia relies on users for fixing errors. Wikipedia started with almost no contribution restrictions, and, as it grew, it developed a few restricted privileges to facilitate administration. It primarily relies on civility and transparency to govern contributors. Any entry can be changed if a contributor thinks that changes are warranted. As stated by a long-time editor who tested a number of articles, "An outsider makes one edit to add a chunk of information, then insiders make several edits tweaking and reformatting it. In addition, insiders rack up thousands of edits doing things like changing the name of a category across the entire site—the kind of thing only insiders deeply care about. As a result, insiders account for the vast majority of the edits. But it's the outsiders who provide nearly all of the content" (Swartz, 2006).

Wikipedia contains many articles that do not differ markedly from those in a printed encyclopedia, such as entries devoted to basic history or science. It also has many entries for general topics in geography and politics. Yet many Wikipedia entries do not neatly fit into a single category, while many are too obscure for attention in a traditional encyclopedia. It faces no limits on the number or size of articles, though a norm developed to keep articles under six to eight thousand words.

3. The Role of NPOV

The Wikipedia site is organized in a way that presumes all errors will be corrected given enough review. This follows a shared assumption among all major participants: Wikipedia follows Linus' Law, "Given enough eyeballs, all bugs are shallow," which Eric Raymond stated in "The Cathedral and the Bazaar."[6]

---

[6] See Raymond (1998), who was rephrasing Linus Torvald, founder of the open-source operating system, Linux. Torvald's rule No. 8 is: "Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix obvious to someone."

Many participants in open-source communities consider Linus' Law to be a foundational principle. For example, ask an editor for Wikipedia about whether Linus' Law works well, and the answer is likely to emphasize the editing process; it comes back to believing in the power of an open-revision process that enables multiple users to edit any passage. Wikipedia's own page about contributing reads, "Many users of Wikipedia consult the page history[7] of an article in order to assess the number of people who have contributed to the article. An article can be considered more likely to be accurate when it has been edited by many different people."[8] Founder of Wikipedia, Jimmy Wales, reiterated the idea in his public comments: "I think the day will come in the future when people will look at an article in Britannica and say, 'This was written by one person and reviewed by two or three more? That's not sufficient. I need an article that's been reviewed by hundreds of people (National Public Radio, 2005).'"

Wikipedia has policies in place to nurture revisions. First, since its founding, Wikipedia has asked all contributors to aspire to write or edit with a NPOV, representing views fairly and without bias. Conflicting opinions are supposed to be presented alongside one another, not asserted in a way that is meant to be convincing. This sometimes is boiled down to the principle to "assert facts, including facts about opinions—but do not assert the opinions themselves."

The cost of representing additional viewpoints is low, so the judgment of the editors creates the primary limit on multiple viewpoints. According to Wales, "If a viewpoint is held by an extremely small (or vastly limited) minority, it does not belong in Wikipedia regardless of whether it is true or not and regardless of whether you can prove it or not."[9]

The limit on minority viewpoints sets up a tension between a narrow and broad interpretation of Linus' Law. If these minority views are held to a minimum, Linus' Law merely requires enough "eyeballs" to operate – with enough contributors all errors will be found. If sampling from many minority views helps achieve a representation of all points of view, then not every contributor can edit all aspects of a

---

[7] Page histories allow a reader to trace the history of edits in reverse chronological order.
[8] http://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia (accessed August 2015).
[9] http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view (accessed August 2015).

controversial or subjective topic. It is possible that Wikipedia must sample from a variety of contributors to achieve NPOV. Indeed, in some articles one or two people have done a substantial fraction of the revisions, while in others the revisions are more evenly distributed across contributors. While there is no bright line between narrow and broad, this tension will inform one aspect of the statistical test about the roles of the total number of contributors and the dispersion of their contributions below.

Verifiability is the second aspiration for contributors. Any reader must be able to check an article's contents and verify against reliable sources. Editors have to be able to cite these sources in their articles and provide links if possible. Editors understand that verifiability is not equivalent to truth; the editor is not responsible for determining whether the information in a newspaper article he or she cites is true, as long as the newspaper is a reliable, peer-reviewed source.

Finally, contributors are asked not to include original research in their contributions. All material must have been previously published by a reputable source. Alternatively, a reasonable adult should understand the concept (i.e., a "vegetable" does not need to be published by a reliable source to be permitted as an article in Wikipedia). This policy was put in place in order to avoid a "novel narrative or historical interpretation" of a subject.[10]

Enforcing these policies and aspirations created many challenges. Over time, the site has adopted a design that makes it simple for contributors to monitor each other. Editors and contributors can subscribe to follow (or "watchlist") articles to check if they have been changed. "Being very transparent encourages good behavior," Wales said (Hyatt, 2006). Furthermore, "Everything is very carefully monitored by a core community who is constantly watching the site, constantly discussing, reviewing changes that are coming in . . . If [a user] is something of an outsider to the community, his changes when they come in will be noticed as, oh, well, this is somebody we don't know and we'll check it over and if it seems fine, it'll stand. Otherwise, it can be removed very quickly (National Public Radio, 2005)."

---

[10] Jimmy Wales, private correspondence, August 28, 2006.

Enforcing NPOV has become the focal point for discussion by those constructing entries in Wikipedia. Many of the back-channel conversations on Wikipedia-dedicated Internet Relay Chat (IRC) concern whether particular passages reflect this principle. In general, the vast majority of entries settles on approaches that the wide community of contributors agrees to, either because such agreements reflect a consensus or because those with minority opinions got the passage they wanted or a dissident gave up.

Could a NPOV ever exist in any of the most controversial topics? Wikipedia's editors point to the triumph of civility on even the most controversial topics, arguing that the results display a more neutral view than any printed entry. They argue that the process takes multiple views into account, achieving something printed encyclopedias do not do as well by relying on a single author.

4. Hypotheses

This study develops a statistical approach for measuring NPOV in the context of Wikipedia articles. That research goal requires translating the collective activities of many contributors, as well as the beliefs of Wikipedia's editors, into testable propositions. This study uses classical statistical approaches, employing a narrow or broad interpretation of Linus' Law, which will constitute the null.

We presume an article is the unit of observation, both at any point in time and over time. Although there are mild exceptions to the constant identity of an article—because some articles are merged or eliminated, etc.—this is a good working assumption for the vast majority of articles. Wikipedia facilitates this approach by assigning numerical identities to articles and maintaining prodigious histories of edits, which helps identify when contributors create new articles and alter (even minor) aspects of existing articles. This also makes it possible to measure the variance in the ages of articles and their conditions over time.

As with other studies of media bias, this study posits that there exists a unidimensional yardstick for measuring neutrality bias. Call this aspect of an article, Y, where Y is a real number that measures the article's political slant. As normalization, let zero be neutral, and loosely speaking, negative is Democrat while positive is Republican. Cardinal numbers have meaning along this yardstick, with larger numbers

denoting more extreme values. Such a yardstick provides two related but somewhat different definitions for neutral/not neutral. One notion is "slant," namely, comparing $Y = 0$ to negative or positive numbers. Another notion compares "NPOV" to "bias," namely, comparing $Y = 0$ to Abs(Y), the absolute value of the slant. The first definition leads to a "slant index" and measures the size of bias *and* its direction. The second definition leads to "bias size" and measures *only* the size of bias.

These measures have intuitive appeal. As it turns out, for example, the data below will show that Wikipedia contains many articles about foreign policy. Some of these articles cover topics, such as the relationship of the US with another country, with little measurable slant and bias. In contrast, some topics (e.g., the Iraq War) tend to have longer entries, and have measurable slant and bias.

This model assumes there is an underlying slant and bias, and it is measured with error. With a large sample, we can test whether slant or bias of a specific article is statistically different from zero. Generally speaking, the estimates perform well (see below), and we estimate tight standard errors, so small deviations from zero are consistent with an article being neutral, and large deviations are not.

This study characterizes the statistical relationship between contributions and Y or Abs(Y). One set of predictions arises intuitively from Linus' Law and its role with NPOV—"Given enough eyeballs, all bugs are shallow." It is possible to proceed under the hypothesis that this law captures a feature of the revision process, namely, that revision attenuates bias. Empirically that means thinly edited pages will have a higher likelihood of bias than thickly edited pages. Stated narrowly:

Hypothesis 1: All other things being equal, an article that has attracted more contributions and contributors over its lifetime will be less extreme than one that has attracted fewer contributions and contributors. Less extreme articles will have a level of Y or Abs(Y) closer to zero.

As noted above, Linus' Law allows for a narrow and a broad interpretation. The narrow version focuses solely on the amount of attention and editing an article receives. The broad interpretation adds to this list the dispersion of contributions. The equivalent of the broad interpretation in the open source environment

would be to ask whether the quality of software programs would be higher when only a few programmers make most of the contributions or when a large number of programmers make a similar number of contributions. In our context, as NPOV may require all conflicting opinions to be presented, an article that has attracted a more diverse set of contributors over its lifetime will be less extreme/biased than one that has attracted a small number contributors, each making many contributions. These two interpretations are not exclusive of each other: If we find that articles that receive a larger amount attention from a more dispersed set of contributors are more likely to have smaller biases, then both the narrow and broad interpretations of the Linus' Law will be supported. Based on the broad interpretation, we can have the following hypothesis:

Hypothesis 2: All other things being equal, an article that has attracted a more diverse set of contributors over its lifetime will be less extreme than one that has attracted a small number contributors, each making many contributions. Less extreme articles will have a level of Y or Abs(Y) closer to zero.

These hypotheses are free of historical context, and many variables will try to control for factors related to vintage and year, such as changes to Wikipedia's size. Such controls are necessary because the Wikimedia Foundation has altered the site over time to enable participation from an increasingly larger group of participants. In addition, many contributors have access to improved broadband technologies, which facilitate online activities, so the composition of online readers has dramatically changed over the decade. There is no reason to expect Wikipedia's contributors to favor one or another political persuasion, on average, so our approach is agnostic with respect to party. The test for Linus' Law will allow for more bias or less bias over time, as the number of contributors increases.

5. Data and Summary Statistics

This study's data come from Wikipedia on January 16, 2011. We develop methods to produce a data set that meets these three criteria: (1) it is possible to measure the NPOV; (2) it is possible to measure the editing process; and (3) within a set of articles, each individual article differs from the others in the amount of attention received.

5.1. Assembling a sample

This study employs a process to maximize the likelihood that at least a few of the articles contain some controversial material, or lack objective data that can be easily verified against outside sources. The initial sample of articles focuses on a broad and inclusive definition of US political topics. It examines the latest version of each article in January 2011, selecting all articles with key words "Republican" or "Democrat," resulting in a list of 111,216 articles. Many of these cover countries other than the United States, necessitating further culling.[11] From this set, we obtain a list of 70,668 articles about US politics.

This sample covers an enormous array of topics, including many controversial ones, such as entries on abortion, gun control, civil rights, taxation, and foreign policy. It also includes many articles that lack anything controversial, such as undisputed historical accounts of minor historical political events and biographies of comparatively obscure regional politicians.

Our approach applies the methods and estimates developed by Gentzkow and Shapiro (2010), hereafter G&S, who developed a method for measuring the biases of US newspapers. Related to G&S, we ask whether a given Wikipedia article uses phrases favored more by Republican or by Democratic members of Congress. G&S select 1,000 phrases based on the number of times these phrases appear in the text of the 2005 *Congressional Record,* applying statistical methods to identify words and phrases that separate

---

[11] The words "Democrat" and "Republican" do not appear exclusively in entries about U.S. politics. If a country name shows up in the title or category names, we then check whether the phrase "United States" or "America" shows up in the title or category names. If yes, we keep this article. Otherwise, we search the text for "United States" or "America." We retain articles in which these phrases show up more than three times. This process allows us to keep articles on issues such as "Iraq War," but drop articles related to political parties in non-US countries.

Democratic representatives from Republican representatives, under the model that each group speaks to its respective constituents with a distinctly coded language. For example, G&S find that Democratic representatives are more likely to use words such as "war in Iraq" and "trade deficit," while Republican representatives are more likely use words such as "economic growth," "illegal immigration," and "border security."[12] After offering considerable supporting evidence, G&S estimate the relationship between the use of each phrase and the ideology of newspapers, using those 1,000 phrases to identify whether newspapers tend to use phrases more aligned with Democrats or Republicans. We label the 1,000 words from the G&S lexicon as "code words."

This approach has several strengths. It has been tested on newspapers and has passed many internal validity tests. In addition, as with newspapers, this provides a general yardstick for measuring the bias of articles, and it removes many subjective elements from that yardstick. Moreover, Wikipedia's contributors are unlikely to have targeted these 1,000 words for editing with this yardstick as a goal, though they might have included or excluded these phrases to try to represent their own views or edit another's views. Finally, G&S's dictionary of code words comes from the middle of our sample, which is fortuitous timing. Since no other nearby year is likely to yield different weights for the code words, that saved us from replicating an extremely time-intensive task for another year of the Congressional record.

This benefit comes with several potential limitations when applying the approach to our setting. First, although newspapers contain hundreds or thousands of code words over time, the measure is quite noisy in a setting with few code words, as occurs on many Wikipedia pages. In one interpretation of G&S, a lack of code words directly indicates that an article lacks bias. In another interpretation, it simply means an article's slant cannot be measured, and it signals little except that the slant index is uninformative. The latter interpretation requires correction for statistical selection, which we employ in our analysis. To say it another way, there are two ways for an article to become neutral, by having no code words or by containing code words with opposite and similar values. We can test whether those two are the same by using standard tests

---

[12] See Table I in Gentzkow and Shapiro (2010) for more examples.

for selection to correct for differences between articles with and without code words (and we find no strong evidence of selection).

Second, a few code words (e.g., "African American" and "illegal immigration") have unusual values for their slant, and in light of the many articles with only a few code words, these outliers could have an inordinate influence on all results. To mitigate their effect, we reset the parameter values for each extreme phrase, namely, the nine most Democrat-leaning phrases and nine most Republican-leaning phrases. We make the value for these phrases equal to the tenth-most left-leaning and tenth-most right-leaning phrase, respectively.

Third, as our approach examines article slant conditional on the topic of the article, our results might be sensitive to the inclusion of different types of articles. For example, articles whose titles contain code words are more likely to exhibit slant because these code words will inevitably be used many times in these articles. To ensure these articles are not driving our results, we identify all articles whose titles contain code words, and exclude these articles from the analysis. Out of 70,668 articles, 1,248 articles (1.8%) are excluded from our analysis. Hence, our final analysis sample contains 69,420 articles.

Finally, the G&S approach identifies not only the choice of phrasing when there are multiple ways of describing the same concept (e.g., "death tax" vs. "estate tax"), but also the choice of topics (e.g., some newspapers choose to run more articles about illegal immigration), while in our study we focus on the former, i.e., the choice of phrasing conditional on the topic. Some phrases in G&S (e.g., "Saddam Hussein," "World Trade Organization," and "Endangered Species Act") do not have natural variations that give no slant or the opposing slant. When such phrases are used, they usually present choice of content rather than actual slant. To ensure that these phrases are not driving article slant, we recruit an experienced copy-editor who has both academic and legal background to go through the 1,000 code words to identify whether there are variations in phrasing for the same concept. We then check all variations she provided. In the end, 638 words out of the 1,000 code words are identified to have substitutes. In our empirical analysis, we use these 638 words as code words to measure slants and biases, ignoring bias arising from the rest of the code words.

We then apply the modified G&S method to compute a slant index for each article.[13] In G&S, articles with no code words have a slant index of 0.49, and articles with slant indices below (above) 0.49 are Democrat-leaning (Republican-leaning). For convenience, we center the slant index for articles with no codes at zero by subtracting 0.49 from all slant indices. We can thus compute the bias size of an article directly as the absolute value of its slant index.

Just as there is no definitive way to measure the "true bias" of a newspaper in G&S, there is no definitive way to measure the "true bias" of a Wikipedia article. Rather, this study uses the distinct words of Republicans/Democrats to measure biases and looks for a series of internal consistency and validity tests. In this sense, "unbiased" and "unslanted" means a balanced number of Republican/Democrat words with similar cardinal values. To give a sense of the numerical value, this method would give a value of -0.07 for the New York Times.[14]

An illustration of how revisions shape NPOV could be the English language article on "Afghanistan,"[15] the country which the United States military invaded after the September 11 terrorist attacks, and in which the US military has played a major role until recently. This article begins in 2002, and in the last few years received several hundred revisions a year. It began with a strong Republican slant (almost reaching a maximum of 0.2), and then revisions increasingly gave it a Democrat slant by 2006 (reaching more than -0.4 for sustained periods and twice reaching -0.6 which were quickly reverted). Since then it has gradually acquired less bias and slant, and at the end of our sample it had a small positive (Republican) value for its slant.

5.2. Descriptive statistics

Of the 69,420 articles observed in January 2011, the last period in which we observe them, it is possible to measure the slant and bias for 19,901 articles (28.7%) – a large fraction of articles contains no code

---

[13] We provide the details in the appendix.
[14] These are from the estimates in Gentzkow and Shapiro (2010), adjusted for the recentering of this article.
[15] See http://en.wikipedia.org/wiki/Afghanistan/, accessed August 2015.

words. As it turns out, 4% have more than 10 code words by this final date. This variance is not surprising, given an oversampling on a wide array of political articles. It is also evidence of skewness in attention at Wikipedia and should not come as a surprise to a frequent Wikipedia participant. Wikipedia includes many articles about obscure political events and individuals that engender little or no attention (e.g., the biography of a mayor of almost any major US city). It also contains another group of political articles about controversial topics (e.g., George W. Bush, Barack Obama, the Iraq War, and health-care legislation) that might attract considerable attention.

Table 1 presents the descriptive statistics of the resulting slant index for these 19,901 articles in January 2011. The table also shows these statistics for different categories of topics in that same year. These categories are not mutually exclusive: Articles can have more than one category attached to them. These categories are assigned by editors and contributors, typically early in an article's life, changing very little over time. The table shows the most commonly used categories.

On average, the average slant is smaller than zero (-0.17) and the difference from zero is statistically significant. Therefore, on average, these articles have a Democrat slant. Most categories have a Democrat slant that differs significantly from zero. For example, articles about civil rights tend to have a Democrat slant (-0.26). At the same time, seemingly controversial topics, such as abortion and trade, are centered around zero.

The 69,420 articles have a total of 16,804,188 revisions. As it is computationally infeasible to examine all these revisions, we take each article and divide its revisions into ten revisions of equal length. For articles with less than ten revisions, we keep all revisions. We retain those revisions, even when they contain no code words. This effort results in 637,641 article observations. Of those, 409,517 of these contain no code words, and at least one code word appears in 228,124 observations (35.8%).

Tables 2 and 3 show how the aggregate statistics vary over time.[16] The statistics for 2001 and 2011 are noisier as we have fewer observations than in other years. Panel A of Table 2 shows there has been movement toward NPOV over time: Wikipedia's articles become less slanted, moving from a mean value of -0.57 in 2002 to a mean value of -0.30 in 2003, and moving gradually downward thereafter to -0.15 in 2010. That would be equivalent to the newspapers in G&S with a strong Democrat slant, such as the San Francisco Chronicle, Atlanta Constitution, or the Baltimore Sun.[17] The standard deviation of this slant index remains large, however. The absolute value of the slant, our bias size, has a similar characteristic, starting at 0.58 (in 2002) and 0.40 (in 2003), and eventually declining to 0.28 (in 2010). Once again, the standard deviation of bias size remains large throughout.

Panel B shows a weighted average across the articles, where the weights come from the number of revisions an article receives in a given year.[18] The number of revisions serves as a proxy for the attention an article receives, and it is the best variable we could assemble that is available for all years and all articles.[19] In both columns, the weighted average is generally lower than the unweighted average. The average slant is lower in the weighted averages. The largest slant is -0.44 (in 2002), and it settles to close to -0.11 in 2010. The largest bias is 0.47 (in 2002), settling to around 0.25 in the later years. Panel B, therefore, suggests that some of the slant and bias in Panel A arises because articles receiving less attention tend to be more slanted and biased.

Panel A of Table 3 provides an overview of how slant and bias change with the age of articles. Most of the older articles lean more Democrat. The bias size has a similar characteristic, with the older articles being more extreme in most cases than the younger ones. In both cases, the standard deviations do not vary much over article ages. Panels B and C look at different vintages of articles at distinct ages. Both panels suggest

---

[16] Different versions of the same article can appear in the same year, so there is no reason to observe the same number of articles each year. Moreover, the last revision of an article may not have been in January 2011, so there will not be a version of every article in 2011.

[17] See Gentzkow and Shapiro (2011), pp. 46-47. Note that the data in this paper have been demeaned, while Genzkow and Shapiro include the original estimates.

[18] The weight is the number of revisions plus one. Because the number of revisions per article is very skewed, this procedure differs little from the alternative, weighting these articles by zero.

[19] A more ideal weight, an article's number of views in a year, is available after 2007. It is highly correlated with revisions (> 0.5) across articles when both are available.

that vintages play an important role and that this role is more important than age. The slant and bias are most pronounced for articles born in 2002 and 2003, with lower slant and bias in all subsequent years. This slant declines mildly with age, with the biggest decline resulting from small samples in the last year (an artifact of the data-collection method). The differences between vintages of articles released in 2002 and 2003 and other vintages also persist.[20]

To summarize, the average old political article in Wikipedia leans Democratic. Wikipedia's articles gradually have lost that disproportionate use of Democratic code words, moving to nearly equivalent use of words from both parties, akin to a NPOV, on average. Moreover, the words used are mildly less extreme over time. The number of recent articles far outweighs the number of older articles, so by the last measurement, Wikipedia's articles appear to be centered close to a middle point, on average. Overall, therefore, Tables 2 and 3 give rise to a question: Why did Wikipedia become less biased over time? What factors in the revision process shape the bias, and what factors determine the appearance of the code words themselves?

Table 4 provides an understanding about the skewness in the revision process. The table is organized around the total share of all revisions, where the first column shows the fraction of total revisions, starting with 0.1 and ending with 0.99. The second column shows the smallest number of articles necessary to obtain the fraction of total revisions in the row. For example, it takes seven articles to account for one percent of the revisions. The next column expresses this same number as the share of articles, dividing by the total number of articles, 69,420. The second to last column provides the average age of articles meeting this criterion, denominated in years. The last column provides the average revisions per year for these articles.

The table demonstrates the skewness of the allocation of revisions. Just over 1.1% of the articles account for 30% of the revisions, and just over 3.3% of the articles account for half of the revisions. Less than 14.6% (27.9%) account for 80% (90%) of the revisions. In short, a small percentage of articles get most of the revisions.

---

[20] Weighted averages indicate similar differences between 2002 and 2003 and other vintages, albeit at lower cardinal values. For the sake of brevity, these are not shown.

Some of this skewness reflects the composition of the age of articles, for example, whether the group includes old or newer articles. Older articles are more likely to have received a larger number of revisions merely by virtue of being around long enough to collect them. Yet, we also show that age cannot be the primary explanation for the skewness. The second to last column suggests that age does not matter much at the most interesting tale of the distribution, where articles get the most revisions. The top 5% are 8.25 years old and the top 50% are 7.61 years old. A half a year hardly accounts for the average difference between these in revisions.

The last column provides a better explanation. As the next column shows, revisions per year differ significantly in each group. For example, the difference of the top 5% and top 50% is 1,739 and 479. In other words, some articles simply get more revisions per year than others, which is a symptom of some articles simply attracting more attention from contributors.

The large change in skewness in revisions per year continues at all levels. For example, in the row for those articles receiving 80% of the revisions per year is 205, contrasting with the 479 received by the top 50%. That implies that the additional 7,806 articles (e.g., 10,110 - 2,304) had to have an average of 124 to bring the average down to 205.[21] That is about one quarter of the rate of revisions as the rate of revisions for the top 50%.

To summarize, Table 4 shows one simple reason why Linus' Law will not work with all the articles, namely, a large fraction of articles do not receive much attention or revision. Table 4 also suggests why an econometric approach is required, as there is considerable variance in the number of revisions an article receives.

5.3. Variables

Table 5 presents descriptive statistics for all key variables. Panels A and B include all revisions for which we can measure slant and bias. Panel C include all revisions. Panel A includes our dependent

---

[21] (204.89*10,110-479.28*2,304)/7,806 = 123.90.

variables. *Slant index* varies between -0.61 and 0.62 and has a mean of -0.19. *Bias size* varies between 0 and 0.62 and has a mean of 0.32. We classify the key explanatory variables into three groups. The first group examines a narrow interpretation of Linus' Law, which we label "attention and editing." The second group examines the broad interpretation of Linus' Law, focusing on the dispersion of contributions. The third group measures features of articles that act as controls.

In the first group we have three variables. We expect that more attention and editing lead to more NPOV. We also use *Log(Unique contributors)* to measure the logarithm of the number of unique users who edited this article in the past. Users are identified by their user ids and Internet protocol (IP) addresses. Different (same) IP addresses are counted as different (same) users. We use *Log(Total revisions to date)* to measure the logarithm of the total number of revisions an article had to date. Finally, we use *Log(Pageviews)* to measure the number of page views in a month for this article. Unfortunately, we have data for this variable only after February 2007, when it first began to be collected.

The second group contains two variables. We use *Revisions per contributor*, defined as *Total revisions to date/Unique contributors*, as one measure of the dispersion of contributions. We also use *Herschman-Herfindahl-Index (HHI)*, based off the concentration of *Unique contributors*. If all revisions are edited by one user in the past, *HHI* will be 1 (just as a monopoly in an industry). A small *HHI* index indicates less concentration, or more dispersion.

Many variables are included in the control group. *Log(Total frequency)* measures the logarithm of the number of code words contained in a version of the article. *Log(Words)* measures the logarithm of the number of words in the observed version of an article. *Total frequency* and *Words* are highly correlated, especially for the sample of data in which *Total frequency* > 0, so only one can be used in a regression. Articles are longer mostly because they attract more attention and more editing. Linus' Law would predict that greater *Total frequency* or *Words* leads to more NPOV. However, because slant arises from the sum of codes words, whether two or twenty or in between, *Total frequency* or *Words* also measures whether more code words tends to slant an index as a statistical artifact.

Log(References) measures the logarithm of the number of references in this version. *References per word*, defined as *References/Words*, measures the extent of verification per length of article. A larger number of references should lead to more NPOV under the null. However, references are also easily manipulated and inflated, so the coefficients estimates need to be interpreted cautiously.

We also create dummies to indicate the year in which the article is created. Under the null, the older vintages have had more opportunity for more attention and more editing, so articles with older vintages should have more NPOV. However, the changing composition of participants on Wikipedia and changing knowledge of those participants about NPOV norms could lead to different interpretations. For related reasons we also add year-specific effects as a further control.

Lastly, we create dummies for the categories listed in Table 1 to control for category effects. These controls are for the statistical tendency of some categories to slant in certain directions, or contain a large bias.

Under a narrow interpretation of Linus' Law, only the amount of attention and editing matters. Under a broader interpretation, the dispersion of contributions also matters.

6. Regression Analysis

6.1. Main results

We use the slant index or the bias size of each article as our dependent variable. These variables distinguish between extreme and mild slant, but that gain requires controlling for selection, as many articles do not contain any code words.

Our selection model for the slant index assumes a production function for slant, $Y^* = f(X)$, where X is the list of explanatory variables as discussed in Section 5.3, and Y is measured with error, so $Y = Y^* + u$. The model assumes a function for observing Y, using the same exogenous variables, as in a standard "type-2" Tobit (Amemiya 1985). Hence, in our first stage, we regress the probability of observing any code words in an article on X. Similarly, the selection model for bias size assumes $\ln[Abs(Y^*)] = h(X)$. Once again, this model becomes a standard "type-2" Tobit.

Tables 6 and 7 present the key results.[22] Table 6 presents the estimates for the first stage of the selection model. One estimate includes *Log(Pageviews)* and the other does not, with a corresponding change in the sample size. Results are similar from both cases. We find that having more unique contributors reduces the probability of having code words, while having more revisions works in the opposite direction, increasing the probability of having code words. However, the value of the coefficients suggests a small overall effect: a 10% increase in the number of unique contributors reduces the chance of having code words by 0.2%, and a 10% increase in the number of revisions increases the chance by 0.1%. More attention, as measured by *Log(Pageviews)*, does not appear to affect the probability of having code words.

The results from dispersion also do not point in one direction with regard to selection. These results indicate that a wider community of contributors produces a variety of effects. More revisions per contributor lead to a lower likelihood of having code words, and the effect is small. A one-standard-deviation increase in revisions per contributor leads to change of probability by less than 1 percentage point. Greater concentration among the number of contributors (larger *HHI*) reduces the likelihood of code words. Compared to other variables, *HHI* seems more effective in affecting the chance of having code words. A one-standard-deviation increase can reduce the likelihood of having code words by more than 3 percentage points.

Features of articles do predict the appearance of code words. Longer articles (more words) or articles with more references have more code words. A 10% increase in article length or the number of references leads to a 1.3 percentage points or 0.3 percentage point increase in the likelihood of having code words, respectively. A one-standard-deviation increase in references per word reduces the likelihood by 1.5 percentage points. Vintage effect seems to be important, too: Articles created in 2002 are much more likely to have code words than articles in other years.

The aforementioned ambiguity contrasts with the results for many of the controls. For example, articles on topics such as abortion, civil rights, government, trade, gun control, social security, and tax reform are

---

[22] We omit coefficents of category and year dummies in all regression tables in the paper for the sake of brevity.

more likely to have code words than other articles, all other things equal, with coefficients all above 0.1. Fixing other variables at their means, being related to civil rights increases the probability of having code words by 15.4 percentage points. Some topics are less likely to have any code words, such as drugs, family & children, and infrastructure & technology. The topic, infrastructure & technology, lowers the probability of having code words by 5 percentage points.

Overall, these results suggest that selection cannot be dismissed as a concern for a few controversial topics and for articles that are long, created in 2002, and have a higher concentration of contributors. For most articles, however, it is not likely to be important.

The first two columns of Table 7 present the results for the second stage of the model of the determinants of the slant index. More revisions produce more Democrat bias, as do fewer unique contributors. More attention, as measured by *Log(Pageviews)*, has the same effect. The effect of unique contributors is stronger than the other two factors and hence more attention and editing leads to less Democrat bias. For example, doubling all three measures increases the slant index by about 0.03. Because Wikipedia articles in our sample exhibit Democrat bias on average, the result suggests that more attention and editing helps achieve NPOV.

While having more revisions per contributor does not seem to make a difference to slant index, more concentration among the number of contributors (greater *HHI*) leads to more Republican slant. A one-standard-deviation increase in HHI increases slant index by 0.03.

The features of articles do predict their slant, however. Older vintages lean Democrat in comparison to later vintages, with a difference of over 0.35, which is a big effect, equal to one standard deviation in the dependent variable. That Democrat leaning is especially pronounced in Wikipedia's second year, 2002.

Many of the category controls also take on interesting directions. Abortion articles lean Republican (0.16), civil rights articles lean Democrat (-0.17), and foreign policy articles lean Republican (0.11), while social security articles lean Republican (0.21). These effects are big in light of the size of the standard deviation in the dependent variable.

Overall, the biggest predictors of slant come from the features of articles, such as their topic and vintage. The best predictor of leaning Democrat is the vintage. Consistent with Table 2, older articles lean Democrat, as do articles observed at earlier moments, and the latter tendency diminishes in recent articles. Hence, the best predictor of leaning Republican is a recent observation of an article with a late vintage on a topic such as abortion. The editing process, such as the number of contributors or their spread, does make a difference, but the effect is smaller than article features.

The last two columns of Table 7 present the results for bias size. There is some supporting evidence for the narrow interpretation of Linus' Law. *Pageviews* is significant, but does not matter much, while the other measures of "eyeballs" do matter. While having more revisions increases the movement away from zero, having more participants decreases it. The former matters, but the latter is more important. Holding other variables at their mean values, from Model (3), doubling *unique contributors* results in an 11% decline in the bias, which is substantial, and doubling in *Total revisions to date* results in a 5% increase in the bias.

The results for the broad interpretation of Linus' Law are weak. *Revisions per contributor* is insignificant. *HHI* is negative, which means more dispersion leads to more bias. A one-standard-deviation increase in dispersion leads to an 8.4% increase in bias.

The biggest predictors of bias come from the features of articles, such as their topic and (for some) vintage. The first two vintages of articles display the largest bias. The 2001 and 2011 vintages also display extreme differences, but this is identified by small samples in each case. Some topics, such as civil rights and government, are also more biased, while topics such as abortion, corporations, gun control, social security, and biographies tend to be more neutral.

The results in Tables 6 and 7 suggest that a few features of the revision process do shape slant and bias, such as the number of unique contributors. This is consistent with a narrow interpretation of Linus' Law. None of the evidence is consistent with a broad interpretation, however. In fact, equal level contributions among contributors generate greater bias. This finding suggests that even with a large number of unique contributors, we are more likely to achieve NPOV if we allow some contributors, presumably those with greater experience or knowledge, to contribute more than others.

6.2. Robustness checks[23]

We first conduct several robustness checks by estimating probit models that examine changes to slant and bias size, controlling for the selection effects. The inferences are similar. We also find no qualitative difference in estimating the basic equations on a sample that removed the youngest articles and those with the least active revision rates. The results supporting a narrow interpretation of Linus' Law also are robust to excluding the tests for the broad interpretation, as well as other features of the articles, such as *words, references, and references per word*.

We also check whether our results are driven by the composition of articles. For example, many of the articles may be on politicians, such as current or former members of congress, state senators, and judges. These articles may mechanically appear to have a slant, because they are likely to have direct quotes from these politicians or discuss their political leanings. As a robustness check, we exclude all articles that belong to the biographies category and repeat the analysis. Our results continue to hold.

As a more stringent robustness check, we check whether our results continue to hold beyond political articles. As many of the categories such as abortion, government, and war and peace could contain articles related to politics, we cannot rely on high-level article categories to identify these articles. We first gather a list of words related to politics[24]. We classify each Wikipedia article to be politics-related if any of the sub-categories it associates with contains one of these politics-related words.[25] This procedure classifies 30,263 articles as politics-related. We then drop these articles and repeat the analysis. Our conclusions continue to hold.

---

[23] Results for all robustness checks are unreported and available upon request.

[24] Source: http://www.macmillandictionary.com/us/thesaurus-category/british/General-words-relating-to-politics-and-government and http://www.english-for-students.com/Noun-Words-for-Politics.html (accessed May 2015).

[25] Detailed sub-category information is provided for most Wikipedia articles by contributors. For example, the article on Abraham Lincoln is associated with sub-categories such as 1809 births, 1865 deaths, 19th-century presidents of the United States, American people of English descent, assassinated United States Presidents, deaths by firearm in Washington, D.C. and Illinois lawyers, among others.

We are also worried about reverse endogeneity: as an article becomes more controversial (i.e., more slanted or biased), it is likely to attract more contributors. If this reverse endogeneity is driving our results, we should expect to see a positive correlation between bias size and unique contributors. But our results show a negative correlation. Controversial articles may lead to edit wars among Wikipedia contributors, however. To ensure that edit wars do not drive our results, following the approach used in Xu and Zhang (2013), for each article, we conduct a robustness check by only keeping the first contribution by each contributor. The approach is aggressive because it eliminates not only all future edit wars between a contributor and others, but also additional edits by the same contributor for each article. In total, 250,219 revisions are dropped. We repeat the analysis with the remaining revisions and our results continue to hold.

Finally, the G&S procedure provides estimates of two distinct dimensions of information, political leaning (positive/negative) and its degree (numerical value). That begs questions about whether the use of Democrat/Republican words, by itself, drives our results, or whether the numerical values do. In addition, these numerical values are derived based on an analysis of the 2005 *Congressional Record*. One may be concerned that the relative importance of these code words may change during our study period. To examine these concerns, we take all the code words, label them as Democrat/Republican, and then compute "market shares" for each article. We then regress market share on the same set of predictors. If the numerical values matter a great deal to our coefficient estimates, then this new process would not yield the same results. The results for the coefficient on *Log(Pageviews), Log(Unique contributors), Log(Total revisions to date)* and *Log(HHI)* remain qualitatively similar, while some of the estimates of controls change in mildly uninteresting ways. The overall inference remains the same, that there is some evidence in favor of a narrow interpretation of Linus' Law, but no evidence in favor of its broad interpretation.

7. Discussion and Conclusion

In this study, we empirically examine whether Linus' Law, a principle that many participants in open-source communities consider foundational, shapes content production. We conduct this test on US political articles in Wikipedia, where Linus' Law would face challenges due to the presence of controversial topics and lack of verified and/or lack of objective information.

Our first set of findings pertains to the general characteristics of Wikipedia's slant and bias over time. In broad terms, in its earliest years, Wikipedia's political entries lean Democrat and tend to be biased. Over time, both traits diminish, on average. By the most recent observations, Wikipedia's entries lack much slant and contain less bias than observed earlier.

Our second set of findings, which points toward persistent bias, is inconsistent with Linus' Law. This arises partly from a vintage effect, partly from the skewed attention of contributors, and partly as a result of the topic. Overall this second set of findings is consistent with only a narrow interpretation of Linus' Law (i.e., Hypothesis 1), one that emphasizes the number of contributors or reviewers. No evidence supports a broad interpretation of Linus' Law (i.e., Hypothesis 2).

Our third set of findings points towards a number of empirical patterns inconsistent with Linus' Law shaping most articles. The majority of articles receive little attention, and most articles change only mildly from their initial slant. Altogether, this is a potential indicator of collective intelligence bias.

These findings can be reconciled with the historical facts. The general tendency toward more neutrality in Wikipedia's political articles appears to arise largely not from revision, but from the entry of later vintages of articles with an opposite point of view from earlier articles. Wikipedia achieves something akin to a NPOV across articles, but not necessarily within them.

The study demonstrates a broad approach for estimating the relationship between features of articles and the revision process, based on testing both narrow and broad interpretations of Linus' Law. We consistently find evidence for a narrow interpretation, at best, and we find no support for a broad interpretation. This result has implications for measuring production involving a large number of contributors, suggesting it is possible to find regularities (around the number of contributors), but that this

effect should not be regarded as the sole determinant of neutrality. The result also shows that an effort to enforce equal contributions from each contributor is likely to increase bias.

Many organizations believe that they could improve their knowledge production by taking advantage of the wisdom of crowds. We show that in the case of Wikipedia, there are aspects such as NPOV that the crowd does not help achieve successfully. The results suggest that organizations, particularly those developing open content for educational purposes, should devote extra efforts to check bias when crowds produce controversial content.

A natural question for future research is whether an alternative production model could perform better in this dimension. For example, would a production model centered on experts, such as the one employed by Encyclopedia Britannica, do a better job than the one used by Wikipedia at achieving NPOV?

Some readers may not conclude that Linus' Law fails to hold, but, rather, that we did not measure the revision process with a proper set of statistics. As with any econometric research, we do not consider our research to be definitive. In that light, these results motivate a number of potential questions. For example, how frequently do articles with distinct biases link to one another? What factors shape the entry of new articles, particularly articles with bias? Does Linus' Law become weaker in practice due to feedback between an article's existing biases and the biases of the revisions it attracts? Hence, we hope that our attempt to measure Linus' Law and NPOV, and to formulate an approach using classical statistical methods, motivates further work on the operation of key principles and their operation within open-content production.

References

Amemiya T (1985) *Advanced Econometrics* (Harvard University Press, Cambridge, MA).
Bernhardt D, Krasa S, Polborn M (2008) Political polarization and the electoral effects of media bias. *Journal of Public Economics* 92(5-6):1092–1104.
Besley T, Prat A (2006) Handcuffs for the grabbing hand? Media capture and government accountability. *American Economic Review* 96(3): 720–736.
Carr N (2008) *The Big Switch: Rewiring the World, from Edison to Google* (Norton, New York).
Chen P, Yang Y, Pavlou P (2013) Enhancing open innovation contests in online markets. Working paper, Temple University.

Chi E, Kittur A, Pendleton BA, Suh B, Mytkowicz T (2007) Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web* 1(2):19.

DellaVigna S, Kaplan E (2007) The Fox News effect: Media bias and voting. *Quarterly Journal of Economics* 122(3):1187–1234.

Gal–Or E, Grelani T, Yildirim TP (2012) The impact of advertising on media bias. *Journal of Marketing Research* 49(1):92–99.

Gentzkow M, Shapiro JM (2006) Media bias and reputation. *Journal of Political Economy* 114(2):280–316.

Gentzkow M., Shapiro JM (2010) What drives media slant? Evidence from U.S. daily newspapers**.** *Econometrica* 78(1):35–71.

Gentzkow M, Shapiro JM (2011) Ideological segregation online and offline. *Quarterly Journal of Economics* 126(4):1799–1839.

Giles, J (2005) Internet encyclopaedias go head to head. *Nature* 438: 900–901.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31(3):493–520.

Gorbetai A (2011) Aligning collective production with social needs: Evidence from Wikipedia. Working paper, Harvard Business School.

Greenstein S, Devereux M (2009) Wikipedia in the spotlight. Kellogg School of Management Case 5-306-507, Northwestern University.

Greenstein S, Zhu F (2012) Is Wikipedia biased? *American Economic Review Papers and Proceedings* 102(3):343–348.

Groseclose T, Milyo J (2005) A measure of media bias. *Quarterly Journal of Economics* 120(4):1191–1237.

Haefliger S, von Krogh G, Spaeth S (2008) Code reuse in open source software. *Management Science* 54(1):180–193.

Halfaker A, Kittur A, Kraut R, Riedl J (2009) A jury of your peers: Quality, experience and ownership in Wikipedia. Conference paper at *Proceedings of the 5th International Symposium on Wikis and Open Collaboration,* October 25-27, Orlando, FL.

Hann I, Roberts J, Slaughter S (2004) Why developers participate in open source software projects: An empirical investigation. *ICIS 2004 Proceedings,* paper 66.

Hann I, Roberts J, Slaughter S (2013) All are not equal: An examination of the economic returns to different forms of participation in open source software communities. *Information Systems Research* 24(3):520–538.

Hyatt J (2006) The wonder of Wikipedia: How to motivate—and control—an army of 30,000 volunteer workers. *Fortune,* June 12.

Kane GC, Ransbotham S (2013a) It's a network, not an encyclopedia: A social network perspective on Wikipedia collaboration. *Academy of Management Annual Meeting Proceedings* 2009(1).

Kane GC, Ransbotham S (2013b) Which came first, the users or the content? Understanding contribution dynamics on Wikipedia, Working paper.

Lawrence E, Sides J, Farrell H (2010) Self-segregation or deliberation? Blog readership, participation and polarization in American politics. *Perspectives on Politics* 8(2010):141–157.

Larcinese V, Puglisi R, Snyder JM (2007) Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *Journal of Public Economics* 95(9):1178–1189.

Lakhani K, Fayard A-L, Levina N, Pokrywa SH (2012) OpenIDEO. Harvard Business School Case No. 612-066.

Levine R (2006) New websites seeking profit in wiki model. *New York Times*, September 4.

MacCormack A, Rusnak J, Baldwin CY (2006) Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Science* 52(7):1015–1030.

Mullainathan S, Shleifer A (2005) The market for news. *American Economic Review* 95(4):1031–1053.

National Public Radio (2005) Wikipedia, open source, and the future of the web. *Talk of the Nation,* November 2.

Piskorski MJ, Gorbatai A (2010) Testing Coleman's social norm enforcement mechanism: Evidence from Wikipedia. Working Paper, Harvard Business School.

Price CJ (2003) Interfering owners or meddling advertisers: How network television news correspondents feel About ownership and advertiser influence on news stories. *Journal of Media Economics* 16(3): 175–188.

Rajagopalan MS, Khanna VK, Leiter Y, Stott M, Showalter TN, Dicker AP, Lawrence YR (2011) Patient-oriented cancer information on the Internet: A comparison of Wikipedia and a professionally maintained database. *Journal of Oncology Practice* 7(5):319–23.

Ransbotham S, Kane GC (2011) Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly* 35(3):613–627.

Raymond E (1998) The cathedral and the bazaar. *First Monday*, http://tinyurl.com/bqfy3s, accessed January 2015.

Reuter J, Zitzewitz E (2006) Do ads influence editors? Advertising and bias in the financial media. *Quarterly Journal of Economics* 121(1):197–227.

Singh PV, Tan Y, Youn N (2011) A hidden Markov model of developer learning dynamics in open source software projects. *Information Systems Research* 22(4):790–807.

Susarla A, Oh J, Tan Y (2012) Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research* 23(1):23-41.

Sunstein CR (2001) *Republic.com* (Princeton University Press, Princeton, NJ)

Swartz A (2006) Who writes Wikipedia? http://tinyurl.com/7n84bsh, accessed January 2015.

Wagner C, Majchrzak A (2007) Enabling customer centricity using wikis and the wiki way. *Journal of Management Information Systems* 23(3):17–43.

Wen W, Forman C, Graham SJH (2013) The impact of intellectual property rights enforcement on open source software project success. *Information Systems Research* 24(4):1131–1146.

Xu S, Zhang XM (2013) Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction. *MIS Quarterly* 37(4):1043–1068.

Yates D, Wagner C, Majchrzak A (2010) Factors affecting shapers of organizational wikis. *Journal of the American Society for Information Science and Technology* 61(3):543–554

Zhang XM, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101(4):1601–1615.

Table 1: Summary Statistics for Slant Index by Category

	No. Obs.	Mean	Std. Err.	One-tailed t-test
All Categories	19,901	-0.17	0.00	***
Abortion	53	0.01	0.04	n.s.
Bios	3,515	-0.09	0.01	***
Budget & Economy	816	-0.07	0.01	***
Civil rights	938	-0.26	0.01	***
Corporations	91	-0.10	0.03	***
Crime	943	-0.11	0.01	***
Drugs	80	-0.10	0.03	***
Education	978	-0.12	0.01	***
Energy & Oil	177	-0.02	0.02	n.s.
Families & Children	290	-0.12	0.02	***
Foreign Policy	1,434	-0.03	0.01	***
Trade	258	0.00	0.01	n.s.
Government	8,747	-0.21	0.00	***
Gun Control	44	-0.06	0.03	*
Health Care	402	-0.10	0.02	***
Homeland Security	375	-0.10	0.01	***
Immigration	258	-0.05	0.02	***
Infrastructure & Technology	781	-0.09	0.01	***
Jobs	498	-0.11	0.01	***
Principles & Values	450	-0.12	0.02	***
Social Security	4	-0.04	0.01	***
Tax Reform	77	-0.14	0.03	***
War & Peace	1,545	-0.08	0.01	***
Welfare & Poverty	263	-0.09	0.02	***

n.s. not significant, *** p<0.01, ** p<0.05, * p<0.1

Table 2: Summary Statistics for Slant Index and Bias Size by Year

Panel A: Unweighted Slant Index and Bias Size Over Time

Year	Slant Index		Bias Size		No. Obs.
	Mean	Std. Dev.	Mean	Std. Dev.	
2001	-0.05	0.32	0.25	0.21	142
2002	-0.57	0.15	0.58	0.10	3,027
2003	-0.30	0.35	0.40	0.24	622
2004	-0.35	0.35	0.43	0.24	3,213
2005	-0.21	0.35	0.33	0.24	6,337
2006	-0.21	0.35	0.33	0.24	19,531
2007	-0.22	0.35	0.33	0.24	25,951
2008	-0.19	0.35	0.31	0.24	29,098
2009	-0.16	0.34	0.30	0.23	31,529
2010	-0.15	0.33	0.28	0.23	34,488
2011	-0.18	0.33	0.30	0.23	9,889

Panel B: Slant Index and Bias Size Over Time Weighted by the Number of Revisions in that Year

Year	Slant Index		Bias Size		No. Obs.
	Mean	Std. Dev.	Mean	Std. Dev.	
2001	-0.14	0.37	0.32	0.23	142
2002	-0.44	0.29	0.47	0.22	3,027
2003	-0.03	0.21	0.17	0.13	622
2004	-0.11	0.26	0.20	0.20	3,213
2005	-0.12	0.28	0.23	0.21	6,337
2006	-0.13	0.30	0.25	0.21	19,531
2007	-0.15	0.31	0.27	0.22	25,951
2008	-0.15	0.31	0.26	0.22	29,098
2009	-0.12	0.31	0.26	0.22	31,529
2010	-0.11	0.31	0.25	0.21	34,488
2011	-0.14	0.30	0.25	0.22	9,889

Table 3: Summary Statistics to Examine Articles' Vintage Effects

Panel A: Slant Index and Bias Size of Wikipedia's Political Articles for Different Article Ages

Age (Year)	Slant Index		Bias Size	
	Mean	Std. Dev.	Mean	Std. Dev.
[0, 1)	-0.14	0.34	0.28	0.23
[1, 2)	-0.12	0.34	0.27	0.23
[2, 3)	-0.16	0.35	0.30	0.24
[3, 4)	-0.20	0.35	0.32	0.24
[4, 5)	-0.23	0.35	0.34	0.24
[5, 6)	-0.26	0.34	0.36	0.24
[6, 7)	-0.27	0.34	0.36	0.24
[7, 8)	-0.32	0.33	0.39	0.23
[8, 9)	-0.34	0.32	0.41	0.23
[9, 10)	-0.14	0.30	0.26	0.21

Panel B: Slant Index of Wikipedia's Political Articles for Different Article Ages and Birth Years

Age (Year)	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
[0, 1)	-0.05	-0.57	-0.29	-0.09	-0.11	-0.10	-0.10	-0.10	-0.10	-0.11	-0.09
[1, 2)	-0.22	-0.55	-0.20	-0.12	-0.12	-0.11	-0.10	-0.07	-0.08	-0.20	.
[2, 3)	-0.11	-0.52	-0.18	-0.12	-0.12	-0.10	-0.09	-0.08	-0.05	.	.
[3, 4)	-0.16	-0.47	-0.17	-0.12	-0.11	-0.10	-0.10	-0.07	.	.	.
[4, 5)	-0.13	-0.46	-0.19	-0.12	-0.11	-0.08	0.02	.	.	.	.
[5, 6)	-0.13	-0.45	-0.17	-0.12	-0.11	-0.14	.	.	.	.	.
[6, 7)	-0.14	-0.42	-0.17	-0.11	-0.08	.	.	.	.	.	.
[7, 8)	-0.15	-0.43	-0.16	-0.01	.	.	.	.	.	.	.
[8, 9)	-0.13	-0.39	-0.05	.	.	.	.	.	.	.	.
[9, 10)	-0.14	-0.12	.	.	.	.	.	.	.	.	.

Panel C: Bias Size of Wikipedia's Political Articles for Different Article Ages and Birth Years

Age (Year)	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
[0, 1)	0.25	0.58	0.38	0.26	0.26	0.26	0.27	0.25	0.26	0.25	0.18
[1, 2)	0.32	0.57	0.31	0.27	0.27	0.27	0.27	0.24	0.25	0.28	.
[2, 3)	0.29	0.55	0.30	0.28	0.27	0.26	0.25	0.24	0.18	.	.
[3, 4)	0.29	0.51	0.29	0.28	0.27	0.26	0.26	0.23	.	.	.
[4, 5)	0.27	0.50	0.31	0.27	0.26	0.25	0.23	.	.	.	.
[5, 6)	0.26	0.49	0.29	0.27	0.26	0.29	.	.	.	.	.
[6, 7)	0.27	0.47	0.29	0.26	0.16	.	.	.	.	.	.
[7, 8)	0.27	0.47	0.29	0.20	.	.	.	.	.	.	.
[8, 9)	0.24	0.44	0.25	.	.	.	.	.	.	.	.
[9, 10)	0.26	0.20	.	.	.	.	.	.	.	.	.

Table 4: The Distribution of Revisions across Articles in January 2011

Total share of all revisions	(n) Number of articles	(n/N) Share of articles	Average Age of articles (Year)	Average Revisions Per Year
0.01	7	0.0001	8.29	2968.85
0.05	59	0.0008	8.25	1738.59
0.10	149	0.0021	8.32	1356.89
0.20	408	0.0059	8.25	999.31
0.30	811	0.0117	8.07	770.90
0.40	1,415	0.0204	7.80	608.66
0.50	2,304	0.0332	7.61	479.28
0.60	3,702	0.0533	7.28	374.33
0.70	5,999	0.0864	6.91	283.93
0.80	10,110	0.1456	6.49	204.89
0.90	19,365	0.2790	5.84	133.79
0.95	30,243	0.4357	5.32	99.15
0.99	50,528	0.7279	4.57	71.98

Definitions
Total share of all revisions = fraction of total revisions.
Number of articles = smallest number of articles to achieve fraction of total revisions.
Share of articles = smaller share of articles to achieve fraction of total revisions.
N = total number of articles, 69,420.
Average age of articles = average age of articles, in years.
Average revisions per year = average revisions per year for articles.

Table 5: Summary Statistics

Panel A: Dependent Variables

Variable	Obs.	Mean	Std. Dev.	Min	Median	Max
Slant index	163,827	-0.19	0.35	-0.61	-0.12	0.62
Bias size	163,827	0.32	0.24	0.00	0.25	0.62

Panel B: Explanatory Variables (Observations for Which Slant Can Be Measured)

Variable	Obs.	Mean	Std. Dev.	Min	Median	Max
Attention and editing						
Log(Unique contributors)	163,827	3.36	1.83	0.00	3.37	9.55
Log(Total revisions to date)	163,827	4.14	1.90	0.00	4.14	10.70
Log(Pageviews)	107,279	6.66	2.40	0.00	6.57	15.64
Dispersion						
Revisions per contributor	163,827	2.79	4.41	1.00	1.94	403.00
HHI	163,827	0.22	0.29	0.00	0.09	1.00
Article features						
Log(Total frequency)	163,827	1.08	0.56	0.69	0.69	6.41
Log(Words)	163,827	7.08	1.07	1.61	7.02	12.20
Log(References)	163,827	1.63	1.61	0.00	1.39	6.86
References per word	163,827	0.01	0.01	0.00	0.00	0.28

Panel C: Explanatory Variables (All Observations)

Variable	Obs.	Mean	Std. Dev.	Min	Median	Max
Attention and editing						
Log(Unique contributors)	637,641	2.41	1.70	0.00	2.30	9.55
Log(Total revisions to date)	637,641	3.06	1.83	0.00	3.00	10.70
Log(Pageviews)	409,517	5.68	2.24	0.00	5.50	15.78
Dispersion						
Revisions per contributor	637,641	2.49	4.02	1.00	1.72	597.00
HHI	637,641	0.33	0.34	0.00	0.18	1.00
Article features						
Log(Total frequency)	637,641	0.28	0.55	0.00	0.00	6.41
Log(Words)	637,641	6.02	1.47	0.00	6.06	12.88
Log(References)	637,641	0.99	1.32	0.00	0.00	6.93
References per word	637,641	0.01	0.02	0.00	0.01	1.00

Table 6: Regression Results for the First Stage Selection Equation (Marginal Effects)

Dependent Variables	(1) Has code words	(2) Has code words
Attention and editing		
Log(Unique contributors)	-0.020***	-0.026***
	[0.003]	[0.002]
Log(Total revisions to date)	0.011***	0.008***
	[0.002]	[0.001]
Log(Pageviews)	-0.0001	
	[0.0004]	
Dispersion		
Revisions per contributor	-0.002***	-0.002***
	[0.000]	[0.000]
HHI	-0.114***	-0.113***
	[0.005]	[0.004]
Article features		
Log(Words)	0.130***	0.134***
	[0.001]	[0.001]
Log(References)	0.033***	0.028***
	[0.001]	[0.001]
References per word	-0.743***	-0.397***
	[0.077]	[0.063]
Year created = 2002	0.251***	0.233***
	[0.008]	[0.006]
Year created = 2003	-0.004	-0.016***
	[0.006]	[0.004]
Year created = 2004	-0.045***	-0.062***
	[0.005]	[0.003]
Year created = 2005	-0.058***	-0.068***
	[0.005]	[0.003]
Year created = 2006	-0.075***	-0.084***
	[0.005]	[0.003]
Year created = 2007	-0.072***	-0.079***
	[0.005]	[0.004]
Year created = 2008	-0.052***	-0.070***
	[0.006]	[0.004]
Year created = 2009	-0.014**	-0.042***
	[0.006]	[0.004]
Year created = 2010	-0.019***	-0.049***
	[0.007]	[0.005]
Year created = 2011	-0.057***	-0.084***
	[0.014]	[0.010]
Category Dummies	Included	Included
Year Dummies	Included	Included
Observations	409,299	637,243

Year created = 2001, Category = Others, and Year = 2001 are used as benchmark groups. Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Table 7: Regression Results for Slant Index and Bias Size

Dependent Variables	(1) Slant index	(2) Slant index	(3) Bias size	(4) Bias size
Attention and editing				
Log(Unique contributors)	0.050***	0.058***	-0.035***	-0.038***
	[0.004]	[0.003]	[0.002]	[0.002]
Log(Total revisions to date)	-0.021***	-0.021***	0.013***	0.012***
	[0.003]	[0.002]	[0.002]	[0.002]
Log(Pageviews)	-0.001*		0.002***	
	[0.001]		[0.000]	
Dispersion				
Revisions per contributor	0.001*	0.0004*	-0.0003	-0.0002
	[0.000]	[0.0002]	[0.0002]	[0.0002]
HHI	0.101***	0.128***	-0.079***	-0.090***
	[0.008]	[0.005]	[0.005]	[0.004]
Article features				
Log(Total frequency)	-0.018***	-0.016***	-0.043***	-0.047***
	[0.002]	[0.002]	[0.001]	[0.001]
Log(References)	0.025***	0.011***	-0.014***	-0.006***
	[0.001]	[0.001]	[0.001]	[0.001]
References per word	-1.068***	-0.391***	0.605***	0.229***
	[0.131]	[0.114]	[0.090]	[0.078]
Year created = 2002	-0.174***	-0.183***	0.107***	0.108***
	[0.006]	[0.005]	[0.004]	[0.003]
Year created = 2003	0.040***	0.042***	-0.023***	-0.031***
	[0.006]	[0.005]	[0.004]	[0.003]
Year created = 2004	0.103***	0.126***	-0.062***	-0.077***
	[0.006]	[0.005]	[0.004]	[0.003]
Year created = 2005	0.121***	0.141***	-0.078***	-0.093***
	[0.007]	[0.005]	[0.005]	[0.003]
Year created = 2006	0.150***	0.177***	-0.091***	-0.112***
	[0.007]	[0.005]	[0.005]	[0.004]
Year created = 2007	0.161***	0.191***	-0.095***	-0.119***
	[0.007]	[0.006]	[0.005]	[0.004]
Year created = 2008	0.171***	0.203***	-0.106***	-0.131***
	[0.007]	[0.006]	[0.005]	[0.004]
Year created = 2009	0.172***	0.205***	-0.101***	-0.127***
	[0.008]	[0.006]	[0.005]	[0.004]
Year created = 2010	0.158***	0.191***	-0.114***	-0.140***
	[0.009]	[0.007]	[0.006]	[0.005]
Year created = 2011	0.196***	0.240***	-0.195***	-0.228***
	[0.026]	[0.025]	[0.018]	[0.017]
Category Dummies	Included	Included	Included	Included
Year Dummies	Included	Included	Included	Included
Inverse Mills ratio	0.043***	0.003	-0.029***	-0.007***
	[0.005]	[0.003]	[0.003]	[0.002]
Observations	409,299	637,243	409,299	637,243

Year created = 2001, Category = Others, and Year = 2001 are used as benchmark groups. Standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Appendix: Procedure for Computing Slant Index

In G&S, for each Congressperson c, they observe their ideology y_c and phrase frequency f_{pc}, the number of times phrase p appears in Congressperson c's speech, for each phrase p. For each phrase p, G&S regress the relative frequency $\overline{f_{pc}}$, where $\overline{f_{pc}} = f_{pc} / \sum_{p \in P} f_{pc}$, on y_c, and obtain the intercept and slope parameters a_p and b_p, for each phrase p. The parameter values, together with the 1,000 phrases identified by G&S, are available at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/26242.

The 1,000 phrases exhibit heterogeneous slants. To mitigate the effect of outlier phrases (e.g., "African American" and "illegal immigration"), we set the parameter values for the 9 most left-leaning phrases and 9 most right-leaning phrases to be the same as the 10th most left- and most right-leaning phrases, respectively. We also drop phrases that have no natural variations that give no slant or the opposing slant and use the remaining 638 phrases in the analysis.

For each Wikipedia article n, we regress $\overline{f_{pn}}$ - a_p, where $\overline{f_{pn}}$ is the relative frequency of phrase p in the article, on b_p for the 638 phrases, and obtain the slope estimate $\overline{Y_n} = \dfrac{\sum_{p \in P} b_p (\overline{f_{pn}} - a_p)}{\sum_{p \in P} b_p^2}$. When an article has none of the 638 phrases, $\overline{Y_n}$ is 0.4975, so we set $Y_n = \overline{Y_n}$ - 0.4975, and use Y_n as our slant index for article n.