Is Wikipedia Biased?

By SHANE GREENSTEIN AND FENG ZHU*

* Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL, 60208, greenstein@kellogg.northwestern.edu and Marshall School of Business, University of Southern California, Los Angeles, CA, 90089-0808, fzhu@marshall.usc.edu. We thank Megan Busse, Michelle Deveraux, Gil Penchina, Scott Stern, Monic Sun, Joel Waldfogel and many seminar participants for their comments. We are responsible for all remaining errors.

As the largest wiki ever and one of the most popular websites in the world, Wikipedia accommodates a skyrocketing number of contributors and readers. At the end of 2011, after approximately a decade of production, Wikipedia supports 3.8 million articles in English and well over twenty million articles in all languages, and it produces and hosts content that four hundreds of millions of readers view each month.[1] Every ranking places Wikipedia as the fifth or sixth most visited website in the United States, behind Google, Facebook, Yahoo, YouTube, and, perhaps, eBay. In most countries with unrestricted and developed Internet sectors Wikipedia ranks among the top ten websites visited by households.[2]

This achievement is astonishing in light of the resources deployed. Wikipedia achieved its size and high profile with minimal staff. Wikipedia is part of a not-for-profit organization. Donations entirely fund the operations. The vast majority of its content comes from volunteer contributors, who sew contributions together with editing and prose.

The predominant outlook of the articles also is astonishing. Since its founding, Wikipedia aspired to present articles that lack biases. A "Neutral Point of View" (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.

This aspiration appears quite plausible in some settings. NPOV should not be difficult to achieve when articles cover uncontroversial topics loaded with objective information that can be verified against many sources. That setting characterizes the vast majority of the Wikipedia articles about established scientific topics, for example. What about topics lacking

[1] http://en.wikipedia.org/wiki/History_of_Wikipedia, accessed December 2011.

[2] http://www.alexa.com, accessed December 2011.

these ideal features? What biases arise in topics where some of the information is controversial, subjective, and unverifiable?

As an illustration of an approach to addressing this question, this study examines the slant within a sample of 28 thousand entries about US political topics. It measures slant at a point in time, and documents its evolution over time, taking an approach in line with the literature examining content bias in media.

The findings show that Wikipedia contains a bias, and the level or direction of bias is not fixed over time. In its earliest years, Wikipedia's political entries lean Democrat on average. Over time, the slant diminishes. This change does not arise primarily from revision of existing articles. Most articles arrive with a slant, and most articles change only mildly from their initial slant. The overall slant changes due to the entry of articles with opposite slants, leading toward neutrality for many topics, not necessarily within specific articles.

The study is interesting for the questions it frames about the processes for aggregation of information and its accumulation in a stock. Such an activity does not follow standard economic models of production—for example, it lacks a regular sequence of activities aimed at producing a pre-specified design for a product or service, and often lacks price signals. In Wikipedia's case, it also lacks the institutions of private property. Wikipedia uses a commons-based approach to aggregate information from a widely dispersed set of contributors, and is oriented toward production of non-proprietary information.

The topic is also interesting in light of the scale and popularity with which these processes arise at Wikipedia, as noted. Moreover, the decade of experience at Wikipedia is well-documented, so it is an ideal setting for scholarship measuring the stock of knowledge created through production of user-generated content.

There have been some studies of Wikipedia, though none examine its biases.[3] As such, examining Wikipedia is a novel topic for the literature on media bias. Scholars have identified many sources of bias in media content. We are closest to studies of the partisan bias of media.[4] We also draw inspiration from studies that stress reader's desire for reinforcement of their prior beliefs.[5]

[3] A range of studies have examined various aspects of information aggregation at Wikipedia (Chi, Kittur, Pendleton, Suh, Mytkowicz, 2007, Ransbotham and Kane, 2011, Gorbatai, 2011, Piskrski and Gorbatai, 2010, Zhang and Zhu, 2011). Some have touched on political topics (Blake, 2006, Brown, 2011), but not the slant of Wikipedia itself.

[4] For example, Larcinese, Puglisi, and Snyder (2007).

[5] For example, Groseclose and Milyo (2005), Mullainathan and Shleifer (2005), Gentzkow and Shapiro (2006), Bernhardt, Krasa, and Polborn (2008), Balan, DeGraba, and Wickelgren (2009) and Gentzkow and Shapiro (2010).

I. NPOV in Wikipedia

The guidelines within Wikipedia state that all articles should aspire to be written or edited with a NPOV. Conflicting opinions are supposed to be presented alongside one another, not asserted in a way that was meant to be convincing. All significant points of view have to be represented in the article. Wikipedia's editors are instructed to "assert facts, including facts about opinions—but do not assert the opinions themselves." [6]

We examine bias, or the lack of NPOV, of Wikipedia articles on political topics. Our focus on political topics maximizes the chances that at least a few of the articles would contain some controversial material with subjective information. Our data come from the January 16, 2011 release of Wikipedia. We use the following procedure to retrieve articles that focus on a broad and inclusive definition of US political topics. We first examine the latest version of each article in January 2011 and select all articles with keywords "republican" or "democrat." We obtain a list of 111,216 articles. We then eliminate these articles that cover countries other than the United States. [7] In the end, we obtain a list of 70,668 articles about US politics.

For each of these articles, we construct a slant index by applying the methods and estimates developed by Gentzkow and Shapiro (2010), hereafter G&S. 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 phrases that separate Democratic representatives from Republican representatives, under the model that each group speaks to its respective constituents with a distinct set of coded language. In brief, we ask whether a given Wikipedia article uses phrases favored more by Republican members or by Democratic members of Congress.

As with G&S's application to newspapers, this approach provides a general statistical yardstick for measuring the slant of articles. However, the application in this study has two differences with the application in G&S. The measure of the slant of newspapers can be compared with other external sources, while no such source exists for Wikipedia. In addition, newspapers contain hundreds or thousands of phrases over time, while many of Wikipedia's articles have few phrases, if any.

---

[6] http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view (accessed July, 2009).

[7] The words "republican" and "democrat" do not appear exclusively in entries about United States 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." If these phrases do not show up more than 3 times in the text, this article is dropped. This process keeps articles such as "Iraq War" but drop articles related to political parties in foreign countries.

Those differences lead to an interpretative ambiguity in our context. With one interpretation, a lack of phrases indicates that an article lacks slant. With another interpretation, lack of phrases means an article's slant cannot be measured. In the results below we generally proceed with the second interpretation, which means all results are conditional on observing any data. It is an open question whether this method observes random articles or selects in ways that accentuate the measured slant. Distinguishing between interpretations is one of the questions raised by this study.

We directly follow the methods outlined by G&S, with a few slight modifications to accommodate a few features of this setting. First, in G&S, articles with no code words will have a slant index of 0.49. Articles with slant indices below (above) 0.49 are left-leaning (right-leaning). For convenience, we center the slant index for articles with no codes at zero. Second, the method applies some trimming to account for outliers. The 1,000 phrases exhibit a few words (e.g., "civil rights" and "illegal immigration") with unusual values for their slant. 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 the same as the tenth most left-leaning phrase and the tenth most right-leaning phrase, respectively.

Of the 70,668 articles observed in January 2011, it is possible to measure the bias for 28,382 articles (40.2%). As it turns out, 3.68% have more than ten phrases by this final date. This variance and skewness is not surprising, given an oversampling on a wide array of political articles. It is evidence of skewed attention at Wikipedia and should not come as a surprise to a frequent participant in Wikipedia. Wikipedia includes many articles about many 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 Bush, Barack Obama, the Iraq War, the Health Care legislation) that potentially attract considerable attention.

Table 1 presents the descriptive statistics organized around topics of the slant index for 28,382 articles in January 2011, the last period we observe them. Articles can and do have more than one topic attached to them. Topics are assigned by editors and contributors, typically quite early in an article's life, changing little over time. The table shows the

most common categories: abortion; budget & economy; civil rights; corporations; crime, drugs; education; energy; families and children; foreign policy; trade, government, gun, health care, homeland security, immigration, infrastructure & technology, employment, value, social security, taxation, war & peace, and welfare & poverty.

TABLE 1— SUMMARY OF STATITICS FOR SLANT

	No. Obs.	Mean	Std. Dev.
All Categories	28,382	-0.09	0.28
Abortion	71	0.02	0.23
Budget & Economy	1,109	-0.02	0.22
Civil rights	1,183	-0.16	0.27
Corporations	121	-0.06	0.24
Crime	1,257	-0.05	0.24
Drugs	105	-0.02	0.20
Education	1,362	-0.05	0.25
Energy	270	-0.02	0.19
Families & Children	405	-0.06	0.24
Foreign Policy	2,094	0.02	0.19
Trade	399	0.06	0.18
Government	11,383	-0.14	0.30
Gun Control	56	-0.03	0.17
Health Care	556	-0.05	0.26
Homeland Security	490	-0.05	0.22
Immigration	372	-0.02	0.22
Infrastructure & Technology	1,143	-0.04	0.24
Jobs	693	-0.05	0.24
Principles & Values	614	-0.05	0.25
Social Security	5	-0.10	0.12
Tax Reform	95	-0.06	0.23
War & Peace	2,292	-0.02	0.21
Welfare & Poverty	323	-0.04	0.22
Bios	4,748	-0.05	0.25

Certain categories of topics tend to differ from zero. For example, when they have a measured slant, articles about civil rights tend to have a Democrat slant (-0.16), while the topic of trade tends to have a Republican slant (0.06). At the same time, many seemingly controversial topics such as foreign policy, war & peace, and abortion are centered at zero. Of course, this table is not meant to be definitive. Rather, it suggests there is considerable variance among topics. Because the standard deviation is often large, it also shows there is considerable variance within topics. Explaining such variance is another open question raised by these findings.

The 70,668 articles in total have 17,270,274 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 of them. This effort results in 647,352 article observations. Of those, 409,363 observations contain no phrases, and we are unable to measure their bias. For 237,989 (36.8%) observations, we have at least one phrase.

TABLE 2— Transition Matrix for slant of first and last article

	Very Negative	Negative	No phrases	Positive	Very Positive
Very Negative	2914	80	1554	38	16
Negative	856	1419	3125	400	68
No phrases	359	98	38891	387	182
Positive	195	335	8967	5167	369
Very Positive	11	22	1788	154	858

Note: The rows are for the latest version and the columns are for the first version of an article.

Table 2 shows the transition between states of slants, taking the earliest and latest observation for each of the 70,668 articles. This table classifies articles into one of five

states: very right, right, no phrases, left, and very left. The cutoff between very right and right is one standard deviation difference from zero, and similarly for left/very left. Though "no phrases" can be interpreted in two ways, as zero or an uninformative, for convenience, it is placed in the center between left and right.

The results show that articles can, and do, change their slants over time as a result of revision, but the changes are rarely dramatic. Very few articles evolve from one extreme to the other—very Democrat and very Republican (only 11 and 16 articles, respectively). Most retain their general direction of bias (generally, more than 60%), and if they transition from one state to another, it is a moderate transition.

TABLE 3 – SLANT OVER TIME

Year	Slant index		No. Obs.
	Mean	Std. Dev.	
2001	0.03	0.24	290
2002	-0.53	0.22	3,276
2003	-0.18	0.33	960
2004	-0.23	0.34	4,571
2005	-0.10	0.30	9,733
2006	-0.11	0.30	28,521
2007	-0.12	0.30	37,465
2008	-0.10	0.29	42,552
2009	-0.08	0.28	46,139
2010	-0.07	0.27	51,210
2011	-0.10	0.27	13,272

Table 3 shows the average slant for the 237,989 articles containing at least one phrase by year.[8] The statistics contain noisiness, particularly in the first and last year, [9] so we are cautious about drawing definitive conclusions. Nonetheless, the table shows there has been movement toward NPOV over time, moving from a mean value of -0.53 in 2002 to a mean value of -0.18 in 2003, and continuing to move gradually downward thereafter to -0.07 in 2010. However, the standard deviation of this slant index remains large, with evidence of a gradual decline, starting in 2002 (0.22), rising in 2003 (0.33), and declining by 2010 (0.27).

We also compute equivalent statistics, weighting them by more or less attention. For the sake of brevity, we do not show the results, but that exercise suggests that some of the slant in Table 3 arises because articles receiving less attention tend to be more slanted. This finding highlights another open question about causality: do more (less) revisions of an article cause the article to contain lower (greater) slant? Potentially causation runs in the other direction too: does less (more) slant cause articles to receive more (less) attention?

Table 4 shows how slant changes with the age of articles. We have 70,636 observations for articles that are less than one year old. We

[8] Note that different versions of the same article can appear in the same year, so there is no reason to observe 27,000 articles each year. There are many articles under the age of one year, but many of these young articles are short, and just getting started. The last revision for an article may not have been in January, 2011, so there will not be a version of every article in 2011.

[9] We see the bias in only 1,292 articles whose age dates them at between at between ages 9 and 10, i.e., a birth in 2001, because this was the first year of Wikipedia.

obtain such a large number because some (very young) articles have multiple revisions with a measured bias, all less than one year old. In that case, all revisions are included. We observe fewer at each successive older age. This supports the conclusion that the trends observed in Table 3 partly result from features of older/younger articles. Most of the older articles lean more Democratic. Two of the three oldest vintages (except the oldest year, which has the smallest sample) lean Democrat in their first year (0.03, -0.53, -0.17), while every other vintage leans Democrat much less strongly in its first year (-0.03, -0.03, -0.05, -0.04, -0.04, -0.04, -0.04).

TABLE 4—SLANT FOR DIFFERENT AGES & YEARS

Age/Year	2001	2002	2003	2004	2005
[0, 1)	0.03	-0.53	-0.17	-0.03	-0.03
[1, 2)	-0.11	-0.51	-0.10	-0.04	-0.05
[2, 3)	0.02	-0.46	-0.09	-0.05	-0.05
[3, 4)	-0.01	-0.39	-0.09	-0.05	-0.05
[4, 5)	-0.02	-0.37	-0.11	-0.06	-0.04
[5, 6)	-0.02	-0.36	-0.10	-0.05	-0.04
[6, 7)	-0.03	-0.33	-0.09	-0.05	-0.08
[7, 8)	-0.04	-0.33	-0.09	0.02	.
[8, 9)	-0.02	-0.29	-0.05	.	.
[9, 10)	-0.04	-0.06	.	.	.

Age/Year	2006	2007	2008	2009	2010
[0, 1)	-0.05	-0.04	-0.04	-0.04	-0.04
[1, 2)	-0.05	-0.05	-0.02	-0.03	-0.08
[2, 3)	-0.04	-0.04	-0.03	-0.05	.
[3, 4)	-0.04	-0.04	-0.06	.	.
[4, 5)	-0.03	0.02	.	.	.
[5, 6)	-0.09	.	.	.	.

Table 4 suggests that the entry of vintages of articles, particularly in Wikipedia's first years, tends to be responsible for differences in the averages that appear in Table 3. The slants are most pronounced for articles born in 2002 and 2003, with lower slants in all subsequent years. These differences decline mildly as articles age, with the biggest decline resulting from small samples in the last year (which is an artifact of the data collection method). In short, the differences between vintages of articles released in 2002 and 2003 and other vintages persist over time.

II. Conclusions and Open Questions

Wikipedia's editor and contributors aspire to generate articles with a neutral point of view. While that goal faces fewer challenges when the information is objective and easily verified, and the topic is uncontroversial, this study examined settings where such conditions are less likely to hold. The study examined a decade of slant in articles about US politics, where some of the articles cover controversial topics, and include inherently subjective and unverifiable information.

The findings show that many of these articles contain bias, and both the level and direction of bias evolves over time. To summarize, the average old political article in Wikipedia leans Democratic. Gradually,

Wikipedia's articles have lost that disproportionate use of Democratic phrases, moving to nearly equivalent use of words from both parties, akin to an NPOV on average. The number of recent articles far outweighs the number of older articles, so, by the last date, Wikipedia's articles appear to be centered close to a middle point on average.

Though the evidence is not definitive about the causes of change, the extant patterns suggest that the general tendency toward more neutrality in Wikipedia's political articles largely does not arise from revision. There is a weak tendency for articles to become less biased over time. Instead, the overall change arises from the entry of later vintages of articles with an opposite point of view from earlier articles.

These results motivate a number of questions about the aggregation of information with a large collection of articles, and about the evolution of the stock of information. For example, how frequently do articles with distinct biases link to one another, cite one another, or maintain distinctly different opinions? What factors shape the entry of new articles, particularly articles with bias? What model explains the feedback from the slant of existing articles to the slant of new entries and revisions to the slant of existing entries? While many studies suggest the distribution of contributions to Wikipedia is quite skewed, how does the distribution of contributions shape slant, and why? Which contributors are most important when it comes to influencing the slant of articles?

This study raises questions about the production of Wikipedia, which generates non-proprietary knowledge with common ownership of aggregated user-generated content. Ultimately, it raises questions about the underlying process, which do not fit existing models of production in which activities produce output following a pre-specified design for the final product or service. The puzzling processes, the scale and importance of the outcome, and the resulting biases in the stock of information, should make user-generated content an object of further economic study.

REFERENCES

Balan, David J., Patrick DeGraba, Abraham L. Wickelgren. 2009. "Ideological Persuasion in the Media." Working paper, Federal Trade Commission.

Bernhardt, Dan, Stefan Krasa, Mattias Polborn. 2008. "Political Polarization and the Electoral Effects of Media Bias." *Journal of Public Economics* 92(5-6): 1092–1104.

Blake, Aaron. 2006. "Wikipedia Site Attempts

to Make Politics Healthier." The Hill, http://www.thehill.com/thehill/export/TheHill/News/Campaign/071106.html, accessed July 2006.

Brown, Adam R. 2011. "Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage." *Political Science & Politics* 44: 339–343

Chi, Ed, Aniket Kittur, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. "Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie." Computer/Human Interaction 2007 Conference.

Gentzkow, Matthew, and Jesse M. Shapiro. 2006. "Media Bias and Reputation." *Journal of Political Economy* 114(2): 280–316.

Gentzkow, Matthew, and Jesse Shapiro. 2010. **"**What Drives Media Slant? Evidence from U.S. Daily Newspapers**."** *Econometrica* 78(1): 35–71.

Gorbatai, Andreea, 2011. "Aligning Collective Production with Social Needs: Evidence from Wikipedia." Working paper, Harvard Business School.

Groseclose, Tim, Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120(4): 1191–1237.

Larcinese, Valentino, Riccardo Puglisi, James M. Snyder. 2007. "Partisan Bias in Economic News: Evidence on the Agenda-Setting Behavior of U.S. Newspapers." *NBER working papers*, National Bureau of Economic Research.

Mullainathan, Sendhil, Andrei Shleifer. 2005. "The Market for News." *American Economic Review* 95(4): 1031–1053.

Piskorski, Mikolaj Jan, and Andreea Gorbatai. 2010. "Testing Coleman's Social Norm Enforcement Mechanism: Evidence from Wikipedia." Harvard Business School Working Paper 11-055.

Ransbotham, Sam and Gerald C. Kane. 2011. "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia." *MIS Quarterly*, 35(3): 613–627

Zhang, Xiaoquan (Michael), and Feng Zhu. 2011. "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review* 101(4): 1601–1615.