Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators

Sean A. Munson, Daniel Xiaodan Zhou, Paul Resnick
School of Information, University of Michigan
1075 Beal Avenue, Ann Arbor, MI 48105 USA
{samunson, mrzhou, presnick}@umich.edu

Abstract

Aggregators rely on votes, and links to select and present subsets of the large quantity of news and opinion items generated each day. Opinion and topic diversity in the output sets can provide individual and societal benefits, but simply selecting the most popular items may not yield as much diversity as is present in the overall pool of votes and links.

In this paper, we define three diversity metrics that address different dimensions of diversity: inclusion, non-alienation, and proportional representation. We then present the Sidelines algorithm – which temporarily suppresses a voter's preferences after a preferred item has been selected – as one approach to increase the diversity of result sets. In comparison to collections of the most popular items, from user votes on Digg.com and links from a panel of political blogs, the Sidelines algorithm increased inclusion while decreasing alienation. For the blog links, a set with known political preferences, we also found that Sidelines improved proportional representation. In an online experiment using blog link data as votes, readers were more likely to find something challenging to their views in the Sidelines result sets. These findings can help build news and opinion aggregators that present users with a broader range of topics and opinions.

Introduction

Observers have raised alarms about increasing political polarization of our society, with opposing groups unable to engage in civil dialogue to find common ground or solutions. Sunstein and others have argued that, as people have more choices about their news sources, they will live in echo chambers (2001). Republicans and Democrats read different newspapers and watch different TV news stations. They read different political books (Krebs 2008). They even live in different places (Bishop 2008). And left-leaning and right-leaning blogs rarely link to each other (Adamic and Glance 2005). If people prefer to avoid hearing challenging views, we may see even greater political fragmentation as people get better tools for filtering the news based on their own reactions and reactions of other people like them.

It is not clear, however, that everyone prefers to be exposed to only reinforcing viewpoints. Stromer-Galley found that participants in online political discussions say that they want and seek out discussions with a diverse range of views (2003). Kelly et al found a diverse exchange of views within political USENET groups (2005): indeed, a good predictor of whether one person was generally liberal was whether the person's respondents were generally conservative in their posts. In a summer 2004 survey, Horrigan et al found that Internet users were more aware of a diverse range of opinions than non-Internet users and that they were not using the Internet to filter out news and opinion that disagreed with their views (2004). In an online experiment, Garrett (2005) found that subjects recruited from the readership of left-wing and right-wing sites were, on average, attracted to news stories that reinforced their viewpoint and showed a mild aversion to clicking on stories that challenged them: once they looked at those stories, however, they tended to spend more time reading them. There appear to be many individuals who seek at least some opinion diversity, though some of them may find it more palatable if also accompanied by some arguments that reinforce their views.

Online news and opinion aggregators such as Digg and Reddit rely on reader votes to select news articles and blog entries to appear on their front pages. They are becoming increasingly popular — Digg, for example, gets more than 35 million visitors each month. Memeorandum selects political news articles and blog entries based in large part on the links among stories: those articles with more incoming links from more popular sources are more likely to be selected. Many aggregators also convene conversations around the articles selected for the front page.

Even if a site selects items based on votes or links from people with diverse views, algorithms based solely on popularity may lead to a tyranny of the majority that effectively suppresses minority viewpoints. That is, even if there is only a slight bias on the input side, there can be a large bias on the output side, a tipping toward the majority viewpoint. For example, if a site has 60 left-leaning voters and 40 right-leaning voters, and each can vote for many articles, then it may be that all the left-leaning articles will get more votes than all the right-leaning articles. Similarly, if a link-following algorithm such as PageRank (Brin and Page 1998) is used on a corpus of blog posts that has 60% left-leaning authors who link predominantly to other left-leaning authors, the left-leaning posts could easily make up

100% of the top ranking articles. If a news aggregator takes no corrective steps, the minority may feel disenfranchised and abandon use of the site, even if they would have been happy to stay and participate in a site that included their viewpoint only 40% of the time. Over time, even people who would prefer to get a mixed selection of news, and to participate in discussions with people who have mixed views, would end up sorting themselves into homogeneous groups.

Diversity Goals

Beyond retaining readers with minority viewpoints, there are several societal reasons why some form of diversity might be valuable. One diversity goal is to make as many people as possible feel that their viewpoint is included in the aggregator's result set People who feel that their view is a minority position and so far unspoken may remain silent to promote social harmony (Rossenburg 1955; Mansbridge 1980): by making people see that their viewpoints are publicly represented in the selected news and opinion items, people may be more likely to articulate their viewpoints in discussion, at the news aggregator site and elsewhere. Moreover, people may be more open to hearing challenging opinions once they feel their own viewpoint is represented (Garrett 2005), so making more people feel included may induce more people to expose themselves to challenging viewpoints.

A second diversity goal is to represent viewpoints in the result set in proportion to their popularity. This could help everyone understand the relative popularity of different viewpoints. There is a natural tendency for people, particularly those in the minority, to think that their own views are more broadly shared than they actually are (Ross et al 1977). Having a better assessment of their true popularity may lead people to accept the legitimacy of disagreeable outcomes in the political sphere, rather than concocting conspiracy theories to explain how the supposed majority will was thwarted.

A third diversity goal is to ensure that everyone is exposed to challenging viewpoints. A long history of experiments has shown that deliberation on an issue with like-minded people leads to polarization: everyone tends to end with more extreme views than they started with (Brown 1985). Awareness of minority views can also lead individuals in the majority to more divergent, out of the box thinking, which can be useful in problem solving (Nemeth and Rogers 1996).

It will be valuable, then, to develop algorithms for news and opinion aggregators that select items that in some way reflect the diversity of opinions of their readers.

Overview of the Paper

In this paper, we present an algorithm, which we call Sidelines, that is intended to increase the diversity of result sets. We compare the results produced by the algorithm to those from a pure popularity selection algorithm, using three diversity metrics, and report on an online experiment that asked people to assess the result sets subjectively.

Data sets

The first domain we explored consisted of user votes on Digg.com. Using Digg's public API, we tracked items and votes in the category "World and Business", which includes political news and opinion during the period October 11, 2008 to November 30, 2008. It had an average of 4,600 new incoming stories and 85,000 diggs (votes from users to stories) from an average of 24,000 users every day. Voting roughly followed a power law – 91% of users voted less than once per day, contributing 28% of the total votes, and 0.7% of users voted more than 10 times per day, contributing 32% of the total.

The second domain consisted of links from blog posts from a collection of 500 political blogs. We selected blogs for this panel from the Wonkosphere directory of political blogs (1,316 blogs). To be included, a blog had to publish the full content of its posts, including markup, as an RSS or Atom feed, had to have posted a blog entry within the previous month, and to have most of its front-page posts be about political topics. This left us with less than our goal of 500 blogs, so we selected others for inclusion by examining the link rolls of blogs already in the sample, until we reached 500.

We coded each of the source blogs based on its political ideology (liberal, independent, or conservative). We consulted both Wonkosphere and PresidentialWatch08, which maintain directories of weblogs classified by political affiliation. In addition, one of the authors read entries from each blog and coded it manually. When the three classifications disagreed, the majority classification prevailed. If a blog was only classified by one of Wonkosphere and PresidentialWatch, and there was disagreement between that source and the reader, we chose the blogger's self-identification (if present) or the third-party (Wonkosphere or PresidentialWatch) assessment. Our panel of blogs contained 257 liberal blogs (52%), 174 conservative blogs (35%), and 63 independent blogs (13%).[1]

We also processed the blog links to reduce document-redundancy (different URLs for the same document). This processing included checking for the same address except for different file extensions (e.g. .html vs. .htm) and collapsing URLs that are identical except for inclusion or exclusion of "www", some parameters such as session IDs, or the default page in a directory (e.g. "index.html"). For 16 mainstream news sites that commonly appeared in the blog links, we also wrote custom rules to match articles that might be found in multiple sections of the website or in different views.

[1] Six additional blogs that we tracked did not post entries during the period October 26 to November 25, 2008. A full list of sources and classifications is available at http://www.smunson.com/bloggregator/sources-icwsm.csv.

In the creation of selection algorithms, we think of blogs as voters, links as votes, and the web pages that they link to as the candidate items for selection. Note that blog entries often link to news stories from mainstream media sites as well as other blog entries and non-blog web pages. Thus, the universe of items was not limited to the entries in the set of 500 blogs. To avoid bootstrapping issues, we collected links to items for two days before the time period during which we generated results. During the period October 24 to November 25, there were a total of 166,503 links to 106,765 distinct items.

Our influence suppression algorithm is based on an intuition that liberal blogs tend to link to the same items as other liberal blogs, but different items than conservative blogs link to. To verify this, we calculated the average Jaccard similarity coefficients (Romesburg 2004) for both intra- and inter- group pairs (Table 1): although the overlap in items linked to was small for a randomly selected pair of blogs, it was more than twice as high when both blogs were liberal or both conservative than when one of each was selected.

	Ideologies of blog pair	Mean Jaccard similarity
Intra-group	Liberal - liberal	0.00306
	Conservative - conservative	0.00302
	Independent - independent	0.00129
Inter-group	Liberal - conservative	0.00132
	Conservative - independent	0.00208
	Liberal - independent	0.00188

Table 1. Mean Jaccard similarity coefficients for inter-group and intra-group pairs of blogs.

Algorithms

We implemented two algorithms. The first, a generalization of approval voting, serves as a comparison point for the second, which is intended to increase diversity. In an approval voting system, each voter votes for as many candidates as desired and the top k vote-getters are selected (Brams and Fishburn 1978). With the blog collections, we treated each blog as a voter and a link from a blog post to an item as a vote from the blog for the item. Because reader interest in news decays with age (hence the term news), we modified the approval voting system to decay each vote (digg or blog link) linearly over time, such that:

$$w = \begin{cases} 1 - t/t_{\max} & \text{for } t < t_{\max} \\ 0 & \text{for } t \geq t_{\max} \end{cases}$$

where w is the vote's weight, t is the time since the link or vote was first detected, and $t_{\max}$ is the time after which the link or vote no longer contributes to the total. The popularity of an item was computed as the sum of the time-weighted votes, Σw. The algorithm then selected the top k items in terms of total time-weighted votes (Algorithm 1). We call this the *pure popularity* algorithm.

The second algorithm, which we call the *Sidelines algorithm,* dynamically suppresses the influence of voters (Algorithm 2). As with the pure popularity algorithm, it incrementally selects the item with the highest total time-weighted votes at each step. After selecting an item, however, anyone who had voted for it is relegated to the sidelines for a few turns; that is, votes from sidelined voters were removed from consideration for the selection of the next few items. Note that the Sidelines algorithm does not take into account the group affiliations of voters or correlations between them. Our goal in this initial study was to investigate whether simply reducing the influence of voters

Algorithm 1 Pure popularity

Param: I {The set of items}
Param: U {The set of users}
Param: V {The users by items matrix of votes; $V_{ui} = 1$ if user u voted for item i, else 0.}
Param: k {Number of items to return}
Param: timeweight() {A function that time-weights votes based on age}

results ← []

for all items *i* **in** *I* **do:**
  i.score ← 0
  for all users *u* **in** *U* **do:**
    i.score ← *i.score* + timeweight(*V*[*u,i*])

results ← first *k* items in *I* sorted by score

Return: *results*

Algorithm 2 Sidelines

Param: I {The set of items}
Param: U {The set of users}
Param: V {The users by items matrix of votes; $V_{ui} = 1$ if user u voted for item i, else 0.}
Param: k {Number of items to return}
Param: timeweight() {A function that time-weights votes based on age}
Param: *turns* {Number of turns to sideline a "successful" voter}

results ← []

{Initially no one is sidelined for any turns}
for all users *u* **in** *U* **do:**
  sidelined[u] ← 0

while length(*results*) < *k* **and** length(*results*) < length(*I*) **do:**
  for all items *i* **in** *I* **do:**
    i.score ← 0
    for all users *u* **in** *U* **do:**
      if *sidelined*[*u*] ≤ *0* **then** *i.score* ← *i.score* + timeweight(*V*[*u,i*])

  {Decrease the sideline turns for each item}
  for all users *u* **in** *U* **do:**
    sidelined[*u*] ← *sidelined*[*u*] – 1

  winner ← item with maximum score
  Remove *winner* from *I*
  Append *winner* to *results*
  for all voters *v* in *V* **do:**
    if *V[u, i] = 1* **then** *sidelined*[*u*] ← *turns*

Return: *results*

whose preferred items had already been selected would noticeably improve the diversity of the result sets.

Diversity Measures

Inclusion / Exclusion. One simple diversity metric measures the proportion of voters who had at least one voted-for item in the result; this is the inclusion score. When computing this metric for any snapshot of results, only voters who had voted for at least one item in the previous 48 hours are included, since votes decay over 48 hours in our time-weighting. The percent who do not have any voted-for items in the result set is the exclusion score. A higher inclusion (and hence lower exclusion) score is one indicator of greater diversity.

Note that we would expect the Sidelines algorithm to include items for at least as many voters as pure popularity, since it gives more weight to votes from voters who do not yet have an item included. Because it is a greedy algorithm that selects the most popular item at each step, however, pathological cases exist where the Sidelines algorithm actually reduces inclusion.

Alienation. A more sophisticated version of exclusion measures the position of the best item in the algorithm's result set rather than just whether any voted-for item is present. The measure generalizes from the Chamberlin-Courant scoring rule for voting systems that select a set of candidates for a committee (1983). The ideal committee is one that minimizes the total alienation of the voters, measured as the sum of all the voter's alienation scores. Finding an ideal committee according to this sum of alienation scores has been shown to be NP-Complete (Procaccia et al 2008).

In our case, we have a sparse set of approval votes from each voter rather than a complete ranking. Moreover, it is natural to think of the result set K=$\{k_1, k_2, \ldots, k_{|K|}\}$ as ordered, since readers will notice the top news stories in a listing before they notice ones lower down. We define the alienation score for user *u* against K as

$$S_{alienation}(\mathrm{K},u) = \begin{cases} \min(i) & \text{where } \mathrm{k} \in \mathrm{K} \cap \mathrm{V_u} \\ |\mathrm{K}|+1 & \text{otherwise} \end{cases}$$

That is, $S_{alienation}(K,u)$ is either the position of the highest item in K that *u* voted for, or |K|+1 if K has no item that *u* voted for. We then define the overall alienation score for K as the sum of individual alienations, normalized by the maximum possible alienation so that values always lie in the interval [1/(|K|+1), 1].

$$S_{alienation}(K) = \frac{\sum_{u \in U} S_{alienation}(K,u)}{(|K|+1)|U|}$$

A lower score indicates improvement on this diversity metric: more people's viewpoints are represented higher up in the result set. As with the simple inclusion/exclusion metric, the Sidelines algorithm will normally decrease the alienation score at least modestly, though pathological cases exist where it could increase alienation.

Proportional representation. A third diversity metric is a generalized notion of proportional representation: we define a divergence scoring function that is minimized when the result set K has votes from different groups in proportion to their representation in the voter population. In the introduction and description of our pilot data, we divided the people and items into Red, Blue, and Purple, with Red people generally voting for or linking to Red items. If the user population were 60% Blue, we suggested that it would be better to select 60% Blue items than to select 100% Blue items, as might occur if we simply take the approval voting outcome.

More generally, suppose that there are groups, G=(g_1, …, $g_{|G|}$), and that each person *u* may have partial affiliation with each group, which we represent by a vector $\mathbf{u}_G = (u_{g_1}, \ldots u_{g_{|G|}})$, with $\sum_{g \in G} u_g = 1$. For a set of users U, we define the representation for the groups as

$$\mathbf{U_G} = \frac{\sum_{u \in U} \mathbf{u_G}}{|U|}$$

Note that, by construction, $\sum_{g \in G} U_g = 1$. That is, the weights express the proportion of total affiliation for each group. In the case where individual affiliations are pure (i.e., each person is affiliated with just one group), the representation vector simply expresses the proportion of users in each of the groups.

Given a set of votes V, for any item *i* we define *i*'s representativeness with respect to the groups' preferences as a vector of weight $\mathbf{i_G}$, with

$$i_g = \frac{\sum_{u \in U} u_g v_{ui}}{\sum_{u \in U} v_{ui}}$$

That is, each vote is weighted by the portion of the voter's affiliation that belongs to the group. The sum of weighted votes divided by the total votes gives the proportion of the total votes that are affiliated with the group. Note that $\sum i_g$=1. In the case where individual affiliations are pure, i_g simply expresses the proportion of all votes for the item that came from users in group *g*.

Then, we define the representativeness vector $\mathbf{K_G}$ on a subset of items K as the mean representativeness over items in the subset:

$$K_g = \frac{\sum_{i \in K} i_g}{|K|}$$

Next we compare the two vectors $\mathbf{U}_G$ and $\mathbf{K}_G$ to compute the amount that the groups' preferences for the subset K diverge from the groups' proportional representation. Interpreting $\mathbf{U}_G$ and $\mathbf{K}_G$ as probability distributions over the groups, we compute the Kullback-Liebler divergence (1951), otherwise known as the conditional entropy:

$$D(\mathbf{U}_G \| \mathbf{K}_G) = \sum_{g \in G} \mathbf{U}_g \log \frac{\mathbf{U}_g}{\mathbf{K}_g}$$

The divergence score is always positive. Lower scores indicate more proportional representation: the mean item representation score is closer to the mean of the user affiliation scores.

Digg.com Evaluation

For the Digg domain, we computed result sets of size 35 for the pure popularity and Sidelines algorithms once per day from November 19-30. For the Sidelines algorithm, the turns parameter was set to 20, meaning that a voter's votes were excluded for the selection of the next 20 items after a voted-for item was selected. For both the Sidelines and pure popularity algorithms, each link counted as one vote when it was first detected and then decayed linearly to 0 over 48 hours (t_{max}).

Averaging over the 12 result snapshots, 65.1% of users who voted for at least one item in the previous 48 hours had at least one voted-for item included in the 35 results. For the Sidelines algorithm, an average of 66.8% of voters had at least one voted-for item selected. The difference is statistically significant (paired t-test, $t(11)=6.05$, $p<0.001$).

The alienation score was also lower (mean 0.476 vs. 0.463). Partly, this results from including a voted-for item for more users. In addition, contingent on having an item selected, voted-for items appeared somewhat earlier in the result sets: the mean position was 6.91 for the Sidelines algorithm and 7.12 for pure popularity ($t(179668)= 5.63$, $p<0.001$).

We do not have a classification of Digg users into opinion groups. Therefore, we were not able to compute a divergence score to measure the proportional representation with respect to opinion groups of selected items vs. the overall population of voters.

Blog Links Evaluation

For the set of 500 blogs, we generated result sets of $k=12$ items using each of the algorithms. In the Sidelines algorithm, a blog sat out the voting for the next *turns*=7 items after an item it linked to was selected. For both the Sidelines and pure popularity algorithms, each link counted as one vote when it was first detected and then decayed linearly to 0 over 48 hours (t_{max}). We generated results snapshots at 6-hour intervals for a period from October 26 to November 25. The Sidelines algorithm achieved a somewhat higher mean inclusion score (0.445) than the pure popularity algorithm (0.419). The difference was statistically significant (paired t-test, $t(120) = 8.701$, $p < 0.001$)

We also computed the alienation score, $S_{alienation}$ for each snapshot (Figure 2). For this calculation, we included only the blogs that had linked to an item in the time window used to generate the snapshot. The mean $S_{alienation}$ for Sidelines result sets was 0.809, and the mean $S_{alienation}$ for pure popularity was 0.796. This difference is statistically significant (paired t-test, $t(120) = 7.864$, $p < 0.001$).

As described previously, we expected that the pure popularity algorithm might tip toward producing very liberally biased results given that the sample of blogs had somewhat more liberal than conservative blogs, and we expected that the Sidelines algorithm would tip less. To evaluate this, we calculated the proportional representation divergence score, $D(\mathbf{U}_G||\mathbf{V}_G)$ for each snapshot (Figure 3).

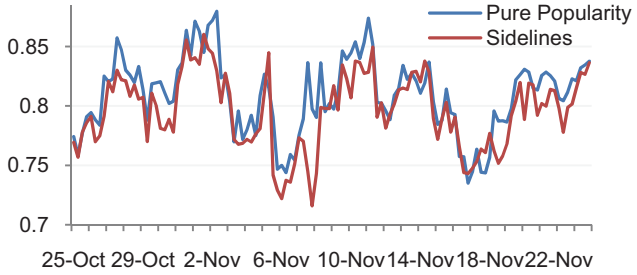

Figure 2. Alienation score for sidelines and pure popularity algorithms on blog data set.

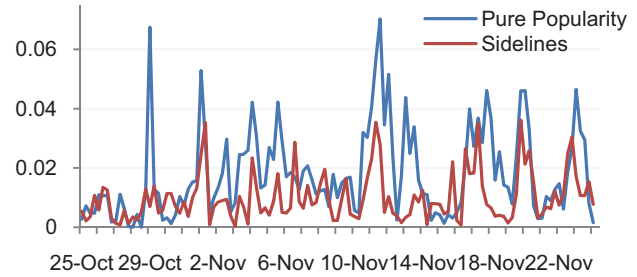

Figure 3. Proportional Representativeness score for sidelines and pure popularity algorithms on blog data set.

The pure popularity algorithm showed some evidence of the expected tipping. While 52% of blogs were classified as liberal (blue), the mean representation of blue opinion among the items selected by the pure popularity algorithm was 61.9% (See $\mathbf{K_B}$ column in Table 2); the mean divergence score $D(\mathbf{U}_G||\mathbf{K}_G)$, which takes into account representation of blue, red, and purple, was 0.018. The Sidelines algorithm showed evidence of tipping as well, but not as severely; the mean representation of blue opinion was 58.6% and the mean divergence score was 0.010. The difference in divergences between the two algorithms was statistically significant (paired t-test, $t(120) = 6.953$, $p < 0.001$). Table 2 summarizes the mean distributions of blue, red, and purple representation for each algorithm.

	$\mathbf{U_B}$	$\mathbf{U_R}$	$\mathbf{U_P}$	$\mathbf{div_{KL}}$
Blog population	*0.520*	*0.352*	*0.128*	-
	$\mathbf{K_B}$	$\mathbf{K_R}$	$\mathbf{K_P}$	$\mathbf{div_{KL}}$
Pure popularity	0.619	0.277	0.103	0.018
Sidelines	0.586	0.313	0.101	0.010

Table 2: Proportional Representation by algorithm

Experiment

To see if readers would notice the difference between the two algorithms, we conducted a web-based experiment from October 27 to December 1, 2008 using the blog links as votes. We recruited subjects using a local list and links posted to the authors' Facebook profiles. Subjects visiting the site were randomly assigned to see the current snapshot of results from one of two algorithms (they were not told which, or indeed that there were two), asked for reactions to each of the items, and then for reactions to the set as a whole. 40 subjects completed the survey; one response was

discarded, as the subject had responded to the questions without actually reading any of the links.

Each subject first was shown links to 12 items, generated by one of the algorithms within the previous 30 minutes. Below each link, we asked subjects to respond to three questions (Figure 4).

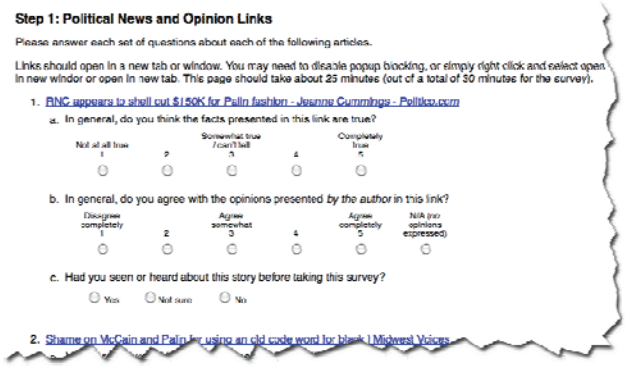

Figure 4. Portion of the online survey asking users to read links and reply to questions.

We then asked readers to respond to the collection as a whole. On a 5-point scale, did the collection seem liberally or conservatively biased? How complete a range of political opinions did they feel the collection included? Did they find opinion-affirming or opinion-challenging items? Did they find something surprising? Subjects were also asked how much they value opinion diversity, topic diversity, opinions that agree with their own, and credibility of facts in news aggregator results.

Finally, we asked subjects about their own political preferences on a 7-point scale from extremely liberal to extremely conservative and on a 7-point scale from Strong Democrat to Strong Republican.

Hypotheses about the Sidelines results. We had several hypotheses related to the diversity goals of the Sidelines algorithm. As one way of measuring opinion diversity, we calculated the variance in how much each subject agreed with the opinions in each article in their result set. A higher variance would indicate a more diverse result set, so we expected that subjects who viewed a list of links generated by the Sidelines algorithm would have a higher variance in how much they agreed with the opinions presented in each article in the collection than subjects who viewed results from the pure popularity algorithm:

H1. Subjects viewing Sidelines algorithm results will report higher variance in their agreement with the opinions in articles presented than subjects viewing the pure popularity algorithm.

As our sample of voting blogs was somewhat liberal, and we anticipated a mostly liberal set of respondents, we hypothesized that the Sidelines algorithm would result in a greater chance of challenging and surprising items.

H2. Subjects viewing a list of links generated by the Sidelines algorithm would be more likely to find something challenging than subjects viewing results from the pure popularity algorithm.

H3. Subjects viewing a list of links generated by the Sidelines algorithm would be more likely to find something surprising than subjects viewing results from the pure popularity algorithm.

Because of the Sidelines algorithm's resistance to tipping, we believed respondents would feel that the Sidelines results represented a more complete and less biased range of items than results from the pure popularity algorithm.

H4. Subjects viewing a list of links generated by the Sidelines algorithm would rate the collection of links as including a more complete range of political opinions than subjects in the pure popularity group.

H5. Subjects viewing a list of links generated by the Sidelines algorithm would rate the bias of the collection as more neutral than subjects in the pure popularity group.

Given Stromer-Galley's findings that people say they seek out diversity, we also expected the Sidelines algorithm would lead to result sets that are more satisfying than the pufre popularity results.

H6. Subjects viewing a list of links generated by the Sidelines algorithm would report being more satisfied with the collection of links than subjects in the control group.

Results. Table 3 summarizes our results from the online experiment.

H1 Opinion Diversity. Though the values are in the expected direction – higher variance in agreement for the Sidelines – the result is not statistically significant.

H2 Challenge. The subjects, who mostly self-identified as Democrats and liberals, were more likely to find something that challenged their opinions in the sidelines result sets; an 89% chance of finding something challenging vs. 50% ($t(35) = 2.83$, $p < .01$).

H3 Surprise. There was no difference in the likelihood of a subject finding a surprising item in the sidelines or pure popularity result sets.

H4 Completeness and H5 Bias. The subjects reported that that the sidelines modification may have delivered slightly more neutral and slightly more complete results sets with regard to range of political opinion, but again, the *p*-values are too small to say this with any certainty.

H6 Satisfaction. There is no apparent difference in subject satisfaction between the two algorithms.

Valued attributes in news aggregator results. We also call attention to the degree to which subjects reported valuing different characteristics of a result set. Subjects indicated that they value diversity of opinions more than agreeable opinions (paired, $t(38)=3.94$, $p < .001$), consistent with Stromer-Galley's findings (2003). Subjects indicated valuing credibility of facts more than either topic diversity (paired, $t(38)=5.97$, $p < .001$) or opinion diversity (paired, $t(38)=6.20$, $p < .001$), and valuing topic diversity more than opinion diversity (paired, $t(38)=2.21$, $p < 0.05$).

Interestingly, readers who viewed results from the Sidelines algorithm reported that they valued diversity in opinions more than the readers who viewed the results from the pure popularity algorithm ($t(35)=2.15$, $p < 0.05$).

Discussion

Free responses indicated that other shortcomings of our aggregator – particularly topic redundancy – drove satisfaction down in both conditions. Unlike many popular news aggregators (e.g. Google News, Memeorandum), we do not cluster similar articles. On days when one particular news story was receiving substantial attention, notably right around the Presidential election, when most of our subjects completed the survey, the list of 12 items presented to survey respondents might contain many links to different coverage of the story; many subjects complained about this in the free response section of the survey. Without the inclusion of these news aggregator features, it is difficult to assess how the Sidelines algorithm affects satisfaction with result sets. Future iterations of our aggregation algorithms – both the baseline we use for comparison and those designed to promote diversity – should cluster related stories by topic to reduce redundancy in the result sets.

While the Sidelines algorithm improved the diversity metrics as compared to pure popularity, effects were not large. Especially with the Digg data, this may be due in part to the large number of voters. For example, in the snapshot taken for November 22, there were 4476 votes for the most popular item. 670 of the 1769 voters for the second most popular item had also voted for the most popular item, but that still left 1099 votes for the second most popular item. We speculate that with so many voters on Digg, even when all the voters for the most popular item are removed, there are many like-minded voters left to vote for the next most popular item. Thus, among the twelve Digg snapshots, the median position in the result set where the pure popularity and Sidelines algorithm first disagreed was not until the fourth item. More importantly, the median position where an item first appeared in the top-35 results with Sidelines that did not appear at all in the pure popularity result set was position 19, with a range of 16-31.

Our user experiment was fairly low-powered. This may have prevented us from obtaining statistically significant results even though some of the user responses were in the anticipated directions. For example, if in the entire population the mean difference in assessments of the collection snapshot's bias was .25 (on a 5 point scale), given the variance in assessments we observed among subjects in our study, the probability of detecting a significant difference (at the p=.05 level) with a sample of 39 subjects is just 14%.

There are two obvious weaknesses in the initial Sidelines algorithm that we used. First, it is more suited to maximizing the number of people who feel their viewpoints are represented in the result set than it is to achieving proportional representation of the different viewpoints of the population, two alternative notions of diversity.

	Pure popularity	Sidelines	p-value
Variance in opinion agreement	1.08	1.38	0.233
Variance in credibility	0.65	0.83	0.210
Credibility (1 not credible, 5 credible)	3.65	3.71	0.708
Overall bias (1 conservative, 5 liberal)	3.40	3.16	0.385
Completeness (1 very incomplete, 5 very complete)	2.50	2.95	0.153
Satisfaction (1 very unsatisfied, 5 very satisfied)	2.60	2.63	0.909
Found something affirming (0 no, 1 yes)	0.94	0.95	0.938
Found something challenging (0 no, 1 yes)	0.50	0.89	** 0.007
Found something surprising (0 no, 1 yes)	0.68	0.68	1.000
Value diversity in opinion (1 not at all, 5 very much)	3.30	4.00	* 0.038
Value diversity in topic (1 not at all, 5 very much)	4.05	4.06	0.982
Value opinion agreement (1 not at all, 5 very much)	2.75	2.79	0.881
Value credibility in facts (1 not at all, 5 very much)	4.45	4.84	0.117
Liberal to conservative (1 extremely liberal, 7 extremely conservative)	2.56	2.06	0.256
Party affiliation (1 strong liberal, 7 strong conservative)	2.55	1.92	0.170

Table 3. Results of online experiment based on political blog links.

Second, because only voters who actually voted for an item sit on the sidelines, and not other people who share their viewpoint, it is not as effective as it could be at getting more viewpoints represented in the final set. We plan to explore different ideas that might improve on the Sidelines algorithm, at least for achieving some types of diversity.

The first approach is to make adjustments to the Sidelines algorithm, changing who is sidelined and for how long. As shown in the pilot study, when we increased the number of rounds users were sidelined before their votes counted again, there was a greater change in the results list for Digg, as compared to the list generated from pure popularity voting. We want to vary this number and find the optimal, which might depend on other parameters such as the size of the set of items, the size of the desired result set, the number of users, and the distribution of votes. To suppress a whole viewpoint, rather than just the voters who voted for an item, we will explore ways to identify those users who share the item's viewpoint and sideline all of them for the selection of the next few items. One way to identify such users would be by clustering: those in the same cluster as those who voted for the just-selected item would be sidelined. Another way would be to use a recommender algorithm: those who are predicted to like the just-selected item would be sidelined.

A second direction for algorithm development is to compute what would be the next selected item, with sidelining, simultaneously for all items, using graph traversal techniques inspired by PageRank (Brin and Page 1998). For example, Jeh and Widom propose a method to compute the pairwise similarities (SimRank) between all pairs

of items by traversing a graph consisting of a node for each pair: the similarity score is interpretable as the expected distance that random surfers starting at the two items would travel before meeting each other if they followed random links in the original graph (2002). One idea for our problem would be to construct a bipartite graph with users and items as nodes. Links from users to items would be the votes of the users, with time-weighted scores. Links from items to users would be included, with weights inversely proportional to the similarity as computed with SimRank. Then, a fixed point of the graph flow could be computed in order to generate a rank score for each of the items; the items with the highest scores would be chosen for the output set.

Despite these limitations and promising directions for future work, it is worth noting one major advantage of the existing Sidelines algorithm: it depends only on the votes for current set of items, and not users' past voting histories or external classifications of the users or items in terms of group affiliations. This characteristic of the Sidelines algorithm is particularly valuable in situations in which more extensive history of user votes or some classification of the voters' or items' viewpoints is not available, such as when users first join a system or when there is a new topic on which previous divisions of people into opinion groups are no longer valid. Simply putting voters on the sidelines for a few turns had a noticeable effect in increasing inclusion and the proportional representation of the result sets.

Conclusion

Opinion diversity, although a desirable feature from a societal standpoint and from the standpoint of at least some individual readers, may not naturally occur when the most popular items are selected. The sidelines algorithm, which suppresses voter influence after a preferred item is selected, is one way to increase diversity. In our experiments, it provided modest increases in diversity according to several different metrics. Perhaps most strikingly, it reduced the tipping toward the majority group even though it operates without any information about the group affinities of people or items, and did so to an extent that was subjectively noticeable. Opportunities remain for research on improved algorithms that take into account additional information, such as group affinities or past voting histories.

Acknowledgements

We wish to thank several anonymous reviewers for helpful comments on an earlier draft of this paper. We also thank Anton Kast at Digg for framing the problem as one of adding diversity heuristics to a pure popularity algorithm. This work has been funded, in part, by a Yahoo Key Technical Challenge grant.

References

Adamic, L., and Glance, N. 2005. The Political Blogosphere and the 2004 US Election: Divided They Blog, *Proceedings of the 3rd international workshop on Link discovery*, pp. 36-43.

Bishop, B. 2008. *The Big Sort: why the clustering of like-minded America is tearing us apart.* New York, New York: Houghton Mifflin Company.

Brams, S. and Fishburn, P 1978. Approval Voting, *American Political Science Review* 72(3), pp. 831-847.

Brin S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems* 30(1), pp. 107-117.

Brown, R. 1985. *Social Psychology, the second edition.* New York: Free Press.

Chamberlin, J. R. and Courant, P. N. 1983. Representative Deliberations and Representative Decisions: Proportional Representation and the Borda Rule, *American Political Science Review* 77(3), pp. 718-733.

Garrett, R. K. 2005. *Exposure to Controversy in an Information Society.* Doctoral Dissertation, University of Michigan. http://hdl.handle.net/2027.42/3974.

Horrigan, J, Garret, K., and Resnick, R. 2004. The Internet and Democratic Debate: Wired Americans hear more points of view. Pew Internet & American Life Project, Washington, DC, 2004.

Jeh, G. and Widom, J.,2002. SimRank: a measure of structural-context similarity, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. Edmonton, Canada, pp. 538-543.

Kelly, J., Smith, M., and Fisher, D. 2005. Opinion Diversity in Online Political Discussion Networks, *Online Deliberation* 2005

Krebs, V. 2008. Political Polarization in Amazon.com Book Purchases. orgnet.com.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency, *Annals of Mathematical Statistics* 22:79-86.

Mansbridge, J.,1980. *Beyond Adversary Democracy.* New York: Basic Books.

Nemeth, C. J, and Rogers J. 1996. Dissent and the search for information. *British Journal of Social Psychology* 35: 67-76.

Procaccia, A. D., Rosenschein, J. S., and Zohar, A. 2008. On the complexity of achieving proportional representation. *Social Choice and Welfare, 30*(3), 353-362.

Romesburg, H. C. 2004. *Cluster Analysis for Researchers.* Lulu Press.

Rosenburg, M. 1955. Some Determinants of Political Apathy. *Public Opinion Quarterly* 18 (Winterly), 349-366.

Ross, L., Greene, D., and House, P. 1977. The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology* 13: pp. 279-301.

Stromer-Galley, J. 2003. Diversity of Political Opinion on the Internet: Users' Perspectives. *Journal of Computer-Mediated Communication* 8(3).

Sunstein, C. 2001. *Republic.com.* Princeton, New Jersey: Princeton University Press.