DebateCAFE v1.0: Incentivizing Articulation and Consideration of Adversarial Arguments

Ken Goldberg
Mo Zhou

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-197
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-197.html

December 6, 2017

DebateCAFE v1.0: Incentivizing Articulation and Consideration of Adversarial Arguments

by Mo Zhou

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Ken Goldberg
Research Advisor

3 Dec 2017

(Date)

* * * * * * *

Anil J. Aswani

Professor Anil Aswani
Second Reader

12-3-17

(Date)

ABSTRACT

In online deliberation systems, self-selection can amplify confirmation bias and cyberpolarization, producing "echo chambers" where initial biases are reinforced rather than explored and resolved. This paper presents DebateCAFE v1.0, a prototype platform that introduces a novel incentive mechanism to encourage participants to articulate persuasive arguments on both sides of an issue. It uses a combination of uncertainty sampling and collaborative filtering to mitigate bias from selective exposure and highlight/rank the persuasive arguments. DebateCAFE v1.0 assigns a score to each participant based on the lower of the Wilson scores of the two arguments entered. To evaluate performance, we used the topic of "Apple (personal privacy) vs. FBI (national security)." Initial results demonstrate the capacity of the system to measure and discourage selective exposure bias without relying on human moderation. We report results from a study with 94 participants who entered 170 arguments on both sides and provided 1754 peer-to-peer ratings on the persuasiveness of the arguments.

DebateCAFE v1.0: Incentivizing Articulation and Consideration of Adversarial Arguments

ABSTRACT

In online deliberation systems, self-selection can amplify confirmation bias and cyberpolarization, producing "echo chambers" where initial biases are reinforced rather than explored and resolved. This paper presents DebateCAFE v1.0, a prototype platform that introduces a novel incentive mechanism to encourage participants to articulate persuasive arguments on both sides of an issue. It uses a combination of uncertainty sampling and collaborative filtering to mitigate bias from selective exposure and highlight/rank the persuasive arguments. DebateCAFE v1.0 assigns a score to each participant based on the lower of the Wilson scores of the two arguments entered. To evaluate performance, we used the topic of "Apple (personal privacy) vs. FBI (national security)." Initial results demonstrate the capacity of the system to measure and discourage selective exposure bias without relying on human moderation. We report results from a study with 94 participants who entered 170 arguments on both sides and provided 1754 peer-to-peer ratings on the persuasiveness of the arguments.

Author Keywords

Interface Design; Social Opinion; Selective Exposure.

ACM Classification Keywords

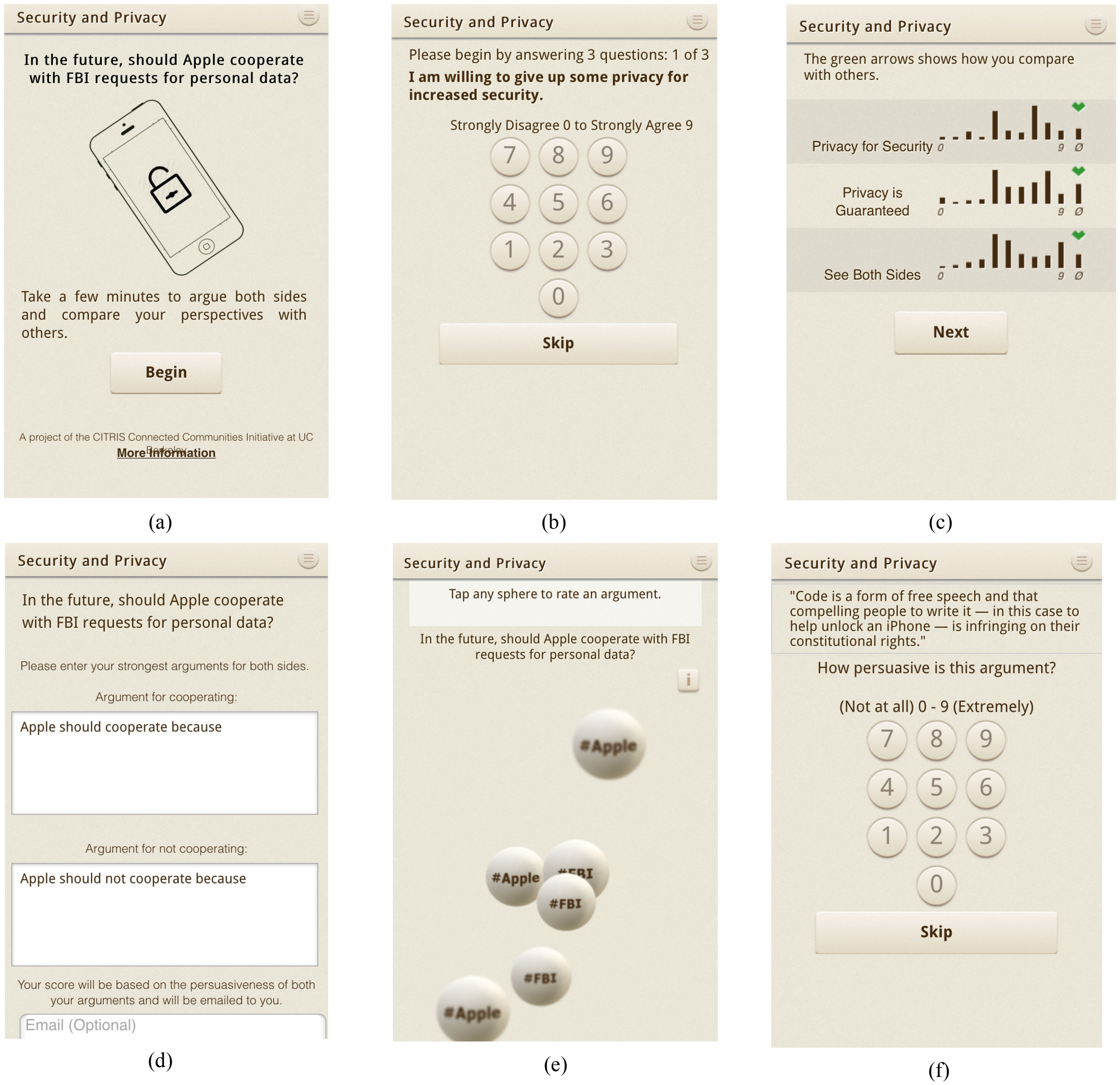
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Online deliberation systems support structured discussion and debate among participants to reach informed decisions that usually involve complex issues and difficult trade-offs [9, 23, 42, 43, 45]. Compared with face-to-face public deliberation, online deliberation has the potential to enable wider participation in civic engagement opportunities and to facilitate group decision-making and problem solving [39].

One potential drawback of such platforms is selective exposure, when participants seek out confirming information and resist changing their opinions. A number of studies acknowledge this issue [1, 7, 10, 14, 21, 28, 29, 34, 37]. Knoblock-Westerwich and Meng show in an experiment that participants are inclined to view articles that are consistent with their own position disregarding the target issue, with 36% more reading time [21]. Salehi et al. found that the decentralizing characteristics of the internet can restrict the discussion to a narrow path in collective intelligence systems [37]. Due to such inherent bias, a small set of initial users with similar positions can dominate discussion, skew the ratings of arguments and eventually reduce the effectiveness of online deliberation. Therefore, greater care is needed when designing such systems to increase awareness of alternative perspectives [5]. Novel approaches have been explored in platforms such as Consider.it [24] and the Deliberatorium [20]. However as the number of participants grows, these frameworks can place a significant burden on human moderators. Novel online deliberation platforms are needed to automate and/or assist the moderation task of filtering out valuable items at scale while mitigating selective exposure bias.

We present the Debate Collaborative Assessment and Feedback Engine (DebateCAFE v1.0), a new platform that seeks to mitigate selective exposure bias with an interface that scales to a large number of participants. DebateCAFE collects quantitative feedback to speculate the initial position of participants, encourages participants to enter convincing arguments on both sides of an issue, and rate the persuasiveness of other participants' arguments. Determining the persuasiveness of the argument is crowdsourced with peer-to-peer ratings, as participants are shown peer arguments (based on an uncertainty sampling algorithm). Participants then receive points for the persuasiveness of their weakest argument. The combination of arguing both sides and a sampling interface avoids allowing a small set of initial participants with similar positions to dominate discussion (a problem highlighted by Salganik [38])



In the future, should Apple cooperate with FBI requests for personal data?

Take a few minutes to argue both sides and compare your perspectives with others.

A project of the CITRIS Connected Communities Initiative at UC Berkeley

(a)

Please begin by answering 3 questions: 1 of 3

I am willing to give up some privacy for increased security.

Strongly Disagree 0 to Strongly Agree 9

(b)

The green arrows shows how you compare with others.

(c)

In the future, should Apple cooperate with FBI requests for personal data?

Please enter your strongest arguments for both sides.

Argument for cooperating:

Apple should cooperate because

Argument for not cooperating:

Apple should not cooperate because

Your score will be based on the persuasiveness of both your arguments and will be emailed to you.

Email (Optional)

(d)

Tap any sphere to rate an argument.

In the future, should Apple cooperate with FBI requests for personal data?

(e)

"Code is a form of free speech and that compelling people to write it — in this case to help unlock an iPhone — is infringing on their constitutional rights."

How persuasive is this argument?

(Not at all) 0 - 9 (Extremely)

(f)

Figure 1: DebateCAFE v1.0 mobile interface screenshots. The Initial Bias Assessment phase asks Likert scale questions to assess initial bias of each participant and displays the results with histograms based on all previous participants. The platform then requests input of two adversarial arguments followed by a graphical display to collect peer-to-peer numerical evaluation of the arguments from other participants.

DebateCAFE uses Collaborative Filtering (CF), an approach widely adopted by companies such as Amazon and Netflix to recommend products. CF allows DebateCAFE to scale to an expanding population of participants and process unstructured textual data on subjective access (i.e., how persuasive is this argument). While collaborative filtering is effective in bringing some structure to lists of items by assigning reputation scores, this should be coupled with incentive mechanisms and an interface design that helps to avoid the biases of selective exposure [5].

This paper explores a case study on the issue of "Apple vs. FBI," inspired by the debate of whether private companies (i.e., Apple Inc.) should cooperate with government (i.e., FBI) requests to have backdoor access to encrypted information technologies. We report data and system performance from a preliminary study with 94 (anonymized university) students who entered 170 arguments on both sides and conducted 1754 peer-to-peer ratings of the persuasiveness of arguments. Results offer valuable insights into platform mechanism design. However, because the participants reflect a narrow demographic (i.e., university students), results may not be reflective of the general population.

The paper is structured as follows: we first evaluate the current state of online deliberation platforms, summarize the design choices that mitigate selective exposure, and survey the theory of incentive to aid the design of our platform. Then we describe the interface and algorithm design of DebateCAFE, in response to the shortcomings of the existing platforms. Finally, we present results from a user study on the "Apple vs. FBI" issue to measure selective exposure and validate the efficacy of the platform.

RELATED WORK

Online Deliberation Platforms

With growing numbers of researchers and government officials recognizing the effectiveness of public deliberation, many online deliberation platforms have been created [6, 8, 13, 16, 17, 20, 24]. The Deliberatorium structures input into issues/problems, ideas/solutions, and for/against arguments [20]. Chilton et al. introduced a collaborative application, Frenzy, to gather the entire program committee for conference session-creation [8]. This crowd-sourced approach significantly reduced the time needed to make a conference program. Kriplean et al. developed a plug-and-play platform called Consider.it, which enables participants to view and articulate arguments on a linear scale of a particular issue. Participants are exposed to pro/con arguments and are encouraged to adopt arguments they find persuasive and articulate pro/con points [24]. Consider.it then requests participants to position themselves on a linear scale to reflect their stand on the issue. Debategraph is another deliberation tool that visually organizes complex debates into graphs, where each node represents an argument and each edge reflects arguments' relationship [3]. Participants are encouraged to provide evidence for each sub-argument. While these platforms are valuable in organizing arguments and collecting feedback, arguments on both sides can be greatly lopsided and the large volume of textual arguments makes it difficult for these systems to harvest valuable insights. DebateCAFE seeks to balance the number of arguments on contradicting viewpoints and uses collaborative filtering to tackle the scale issue. DebateCAFE further introduces a scoring mechanism to incentivize persuasive arguments on both sides.

Selective Exposure

Despite the growing number online deliberation platforms, a major challenge that reduces their effectiveness is selective exposure. Sears and Freedman describe selective exposure as: "People prefer exposure to ideas that agree with their pre-existing opinions" [41]. An inherent characteristic of the internet is the freedom to select from a range of content. However, this freedom enlarges the concern that individuals only view information that aligns with their personal attitudes. Research has shown that people differ in their willingness to be exposed to adverse information, especially political content [35, 40, 44] and that exposure to online content is tailored by algorithms, creating what Pariser refers to as "filter bubbles" that limit exposure to divergent viewpoints [33]. This phenomenon of unbalanced exposure could largely hinder people from making rational and informed decisions and can impede political tolerance [22]. To mitigate selective exposure, Graells-Garrido et al. built a platform that mixes a visual interface and a recommender algorithm to recommend politically diverse profiles to each user and found that an indirect approach in developing systems that reduce user behavior bias can be beneficial [14]. Gao et al. designed a social forum interface that displays controversial social opinions with user reactions to various stances. Results suggest that showing different stances for each topic and providing user reactions to each topic can effectively mitigate selective exposure [11]. However, when the crowd starts to grow, their interface will rely greatly on moderators to categorize new insights, which could quickly become unwieldy. Moreover, directly revealing summarization of stances can be intimidating to those users in the minor group and discourage further participation. Alternatively, DebateCAFE is designed to mitigate selective exposure by implementing an uncertainty-sampling phase to indirectly encourage participants to view and rate a subset of arguments for a pair of contradicting positions, enabling the platform to identify the top arguments even at scale. While making it easy for participants to view and articulate adversarial arguments, the evaluation phase of DebateCAFE also avoids directly exposing participants to opposing arguments by giving participants the freedom to choose which sphere (along with its stance) to click and rate (as shown in Figure 1e).

Collaborative Filtering (CF)

Scaling is another challenge faced by online deliberation platforms. One approach to handle the scale issue is to leverage the crowd by collaborative filtering. Collaborative filtering aggregates subjective ratings provided by humans to assign a numerical reputation to each item [15]. In many CF systems, such as Amazon and Netflix, the reputation of each item is based on the aggregated ratings from a neighborhood of similar items [30]. CF can also be applied globally when the reputation of an item depends on the aggregated ratings from ALL participants, reflecting a common opinion of the crowd.

Most CF systems adopt a list-based presentation, resulting in unbalanced exposure of items, where highly rated items are shown on top [48]. Though the system does not set out to bias any item, the self-selection nature of humans could hinder the presentation of new items due to limited exposure [31]. This can be particularly counterproductive for online deliberation platforms, where arguments on one side could have greater exposure, discouraging participants to articulate and rate arguments with contradicting views. In DebateCAFE v1.0, we balance the exposure of arguments by simultaneously displaying a subset with balanced positions and mixed rankings, permitting the system to present arguments on both sides equally and collect feedback on all arguments.

Theory of Incentives

Selective exposure and biased discussion on existing online deliberation platforms are results of conflicting objectives,

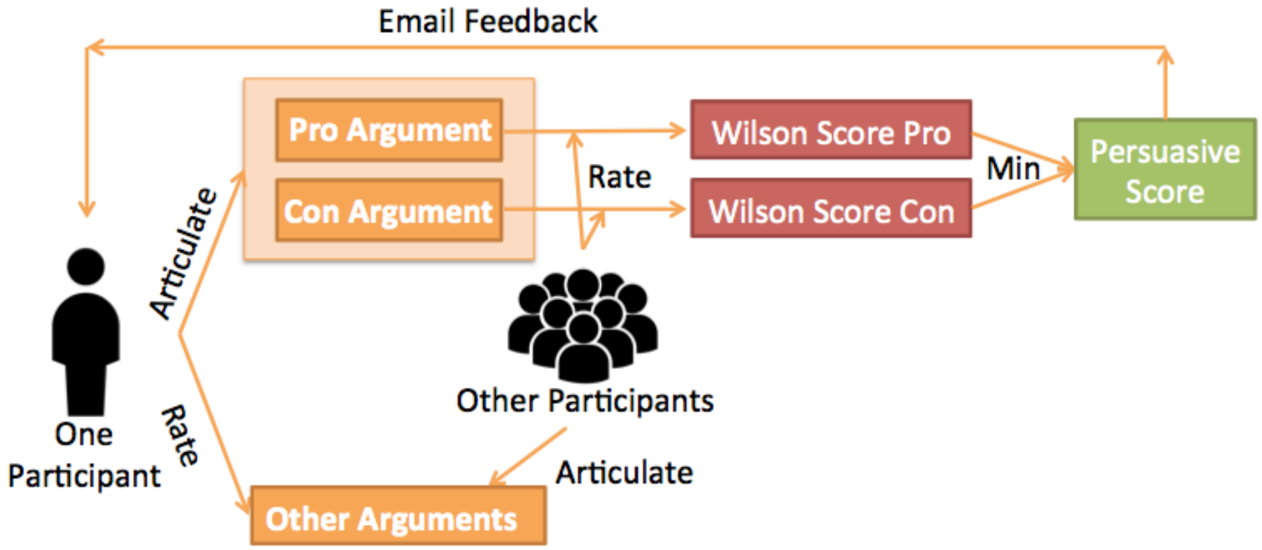


Figure 2: Information flow of DebateCAFE: each participant articulate two arguments for opposing positions of an issue. Other participants rate the persuasiveness of the two arguments. Then, ratings are aggregated using the Wilson metric and the persuasive score for the participant is computed as the lower of the two argument scores and is sent back to the participant via email. In addition, participants are encouraged to rate other arguments on an uncertainty-sampling interface.

where users maximize their persuasiveness on the position they agree with and the platform wishes to optimize the persuasiveness of all positions. The theory of incentives has identified conflicting objectives and decentralized information as two basic factors that lead to inefficient outcome [27]. In particular, the principal-agent model is an abstraction of our use case [11, 18].

In the principal-agent model, the "agent" is able to make decisions on behalf of the impact of the "principal", where both parties (the agent and the principal) have contrary interests and asymmetric information [4]. Oftentimes, it is costly for the agent to perform the activities that are useful to the principal. This model has wide adoption in corporate management and is a core motivation of contract theory. A game theory approach of the problem suggests that an introduction of rules of the game so that the agent interest coincides with that of the principal can be beneficial [36]. DebateCAFE explores this approach to create additional rules to incentivize users to provide strong arguments even for positions they oppose.

PLATFORM DESCRIPTION

In this section, we first describe the user interface of DebateCAFE. Next we introduce the incentive mechanism implemented in the platform to encourage participants to articulate strong arguments for both sides of the issue. At last, we describe an instance of the platform centered on the "Apple vs. FBI" issue.

Interface

DebateCAFE guides participants through three stages: assessment, argument articulation, and peer-to-peer argument evaluation.

Assessment Phase

Participants first assess their current beliefs of the discussed issue by rating three Initial Bias Assessment questions (see *the Apple vs. FBI issue* section for an example). These statements should be able to summarize participants' initial stance. Participants have the option to skip any question they choose not to answer by either pressing the skip button or leaving the response blank (Figure 1b).

Participants are then asked to provide their zip code. We find zip code to be an informative demographic statistic, while not being so intrusive as to hinder further participation in the system. After that, DebateCAFE displays the histograms of the responses to the three IBAs so that the participant learns his/her position on these questions relative to all participants (Figure 1c).

Argument Articulation Phase

In this phase, participants are prompted to formulate their own arguments in response to the central discussion question (Figure 1d). Compared with previous CAFE instances, DebateCAFE is novel in that participants are prompted to enter two arguments: one for each side. The interface strongly encourages, but does not require that a participant fill in both arguments. DebateCAFE also requests that participants supply their email address so we can update them with peer-rating scores on their arguments (explained further in the next section).

Peer-to-Peer Argument Evaluation Phase

Next, participants enter the "discussion space," a 2D visualization where other participants' arguments are represented by spheres arranged across the space (Figure 1e). This discussion space displays 8 argument spheres at a time, 4 “pro” and 4 “con” arguments, and the stance of the argument is labeled on the sphere. Notice the actual arguments are not displayed in this page and participants have the freedom to click on whichever arguments they wish to view. In order to better ensure that all arguments are seen and rated, DebateCAFE prioritizes the display of arguments that have high uncertainty in their evaluation grades. We quantify uncertainty of argument i using the standard error SE_i:

$$SE_i = \frac{SD(R_i)}{\sqrt{N_i}}$$

where R_i are a list of ratings for argument i and N_i is the number of ratings argument i receives.

The spheres are placed in the 2D space according to the first two dimensions of a Principal Component Analysis (PCA) applied to participants’ responses during the assessment phase [47]. (Skipped questions are assigned the mean response rating for that question.) Thus spheres that are closer to the center of the space are provided by more similar participants in terms of IBAs. Participants then click on the spheres in the 2D space to read other participants’ arguments. Finally, using a rating interface similar in design to that used during the assessment phase (Figure 1f), participants evaluate their peers’ arguments on the question “How persuasive is this argument?” using a scale from 0 (Not at all Persuasive) to 9 (Extremely Persuasive).

Persuasive Scoring Mechanism

A key challenge of online deliberation platforms is the conflict of interest between users and the platform, where users optimize their persuasiveness of their position and the platform aim to collect valuable arguments for all positions. To resolve this conflict, DebateCAFE not only provides an uncertainty-sampling interface in the argument evaluation phase, but also introduces a “persuasiveness score” for each participant. We first describe how we compute the score of each argument and then explain the aggregation procedure.

In crowdsourced rating systems, using the average rating for assessment is not robust due to small sample size. Therefore, in DebateCAFE, we calculate argument score with the lower bound of the binomial proportion confidence interval (also called the Wilson Score) [46]. Intuitively, this approach is more robust because it incorporates information about the uncertainty of the score estimate. For example, an argument that receives ratings 10, 0 is ranked lower than one that is rated 5, 5.

Now each participant receives two scores (s_1, s_2) for his/her two arguments on opposing positions. We adopt the insights from the theory of incentive to define the overall persuasiveness of a participant as the minimum of these two scores $\min(s_1, s_2)$. We can view our problem as a principle-agent problem where the platform is the principal and users are agents. We assume that users aim to maximize the persuasiveness of their own position and providing a strong argument for the opposing position lowers their utility. Conversely, the platform tries to optimize argument quality for all positions. Under this setup, we can derive the utility optimization problem of a user in the absence of the persuasive score mechanism as:

$$\begin{aligned} max \quad & U_{Init} = |s_1 - s_2| \qquad (1) \\ s.t. \quad & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned}$$

where v_1, v_2 are the scores of the user’s most persuasive arguments on the two positions.

Without loss of generality, we can assume that this user believes in position 1, then (1) simplifies to:

$$\begin{aligned} max \quad & U_{Init} = s_1 - s_2 \qquad (2) \\ s.t. \quad & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned}$$

Therefore, the optimal solution is $s_1 = v_1$ and $s_2 = 0$, i.e., this user will provide his/her strongest argument for position 1 and provide his/her weakest argument for position 2.

Now we introduce the persuasiveness scoring mechanism and further assume that users gain utility from their overall persuasiveness. Notice also that there is asymmetric information since the platform does not know v_1, v_2. If we let the value of persuasiveness to be λ, then the principal-agent model becomes:

$$\begin{aligned} max \quad & s_1 + s_2 \qquad (3) \\ s.t. \quad & s_1, s_2 = \arg max\ |s_1 - s_2| + \lambda \cdot \min(s_1, s_2) \\ & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned}$$

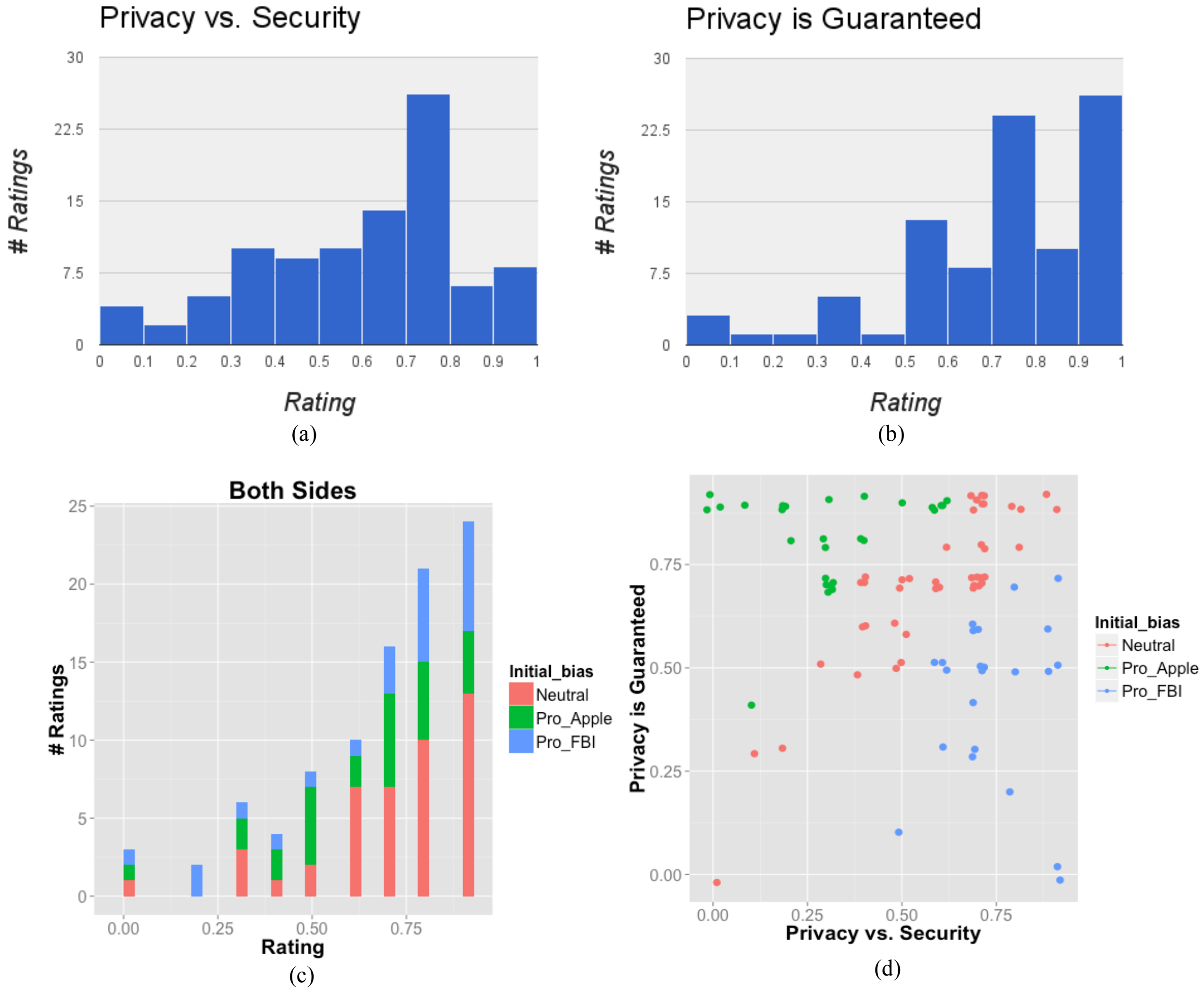
and each user solves the updated utility optimization problem as follows:

$$\begin{aligned} max \quad & U = |s_1 - s_2| + \lambda \cdot \min(s_1, s_2) \qquad (4) \\ s.t. \quad & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned}$$

Let’s again assume this user believes in position 1, then we can simplify (4) to:

$$\begin{aligned} max \quad & U = s_1 - s_2 + \lambda \cdot s_2 \qquad (5) \\ s.t. \quad & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned}$$

Now when $\lambda > 1$, the optimal solution is $s_1 = v_1$ and $s_2 = v_2$, i.e., the user will provide his/her strongest argument for both positions, which aligns with the incentive of the platform. Notice that even though the value of λ is user-specific (as users can have different valuations on being persuasive), it can be enhanced by additional mechanism designs. For instance, the platform can limit the exposure of arguments provided by users with a low persuasiveness score or having a public leaderboard showing the most persuasive users. We plan to explore the impact of these additional designs in the next version of DebateCAFE.



(a)

(b)

(c)

(d)

Figure 3: (a), (b), (c): Histogram of IBA 1-3 respectively. (d): Plot of IBA2 versus IBA1, color coded by the estimated initial position of the participant.

The Apple vs. FBI Issue

To evaluate platform performance, we chose the topic of "Apple vs. FBI." The issue arose when the FBI requested that Apple help investigators gain access to an iPhone used by Syed Rizwan Farook in a December 2015 mass shooting in San Bernardino, CA [2]. Apple refused the request because it would require writing new software to bypass encryption features of the iPhone and would create "the potential to unlock any iPhone in someone's physical possession" [19]. This particular incident attracted public attention and led to debates on social media platforms and related online forums. Starting with this incident, the discussion has expanded to a more general theme of "personal privacy vs. national security". We perceive this issue as a highly complex and controversial topic that may be clarified through deliberation, allowing the public to articulate ideas side-by-side.

For this issue, we encourage participants to rate the following three IBA questions on a 10-point scale from 0 (Strongly Disagree) to 9 (Strongly Agree).

1. *I am willing to give up some privacy for increased security.*
2. *Personal privacy should be guaranteed by the US Constitution.*
3. *There are reasonable arguments on both sides of the security and privacy debate.*

These statements were chosen as they succinctly summarized participants' initial stance (pro-Apple vs. pro-FBI) on digital privacy, as well as their degree of open-mindedness to opposing arguments. In the argument articulation phase, the main discussion question is, "In the future, should Apple cooperate with FBI requests for personal data?" To help bootstrap participant's writing, we pre-populate the textboxes with "Apple should cooperate because", and "Apple should not cooperate because." Furthermore, in the peer-to-peer argument evaluation phase, each sphere is labeled either "Apple" or "FBI" to reflect the stance of the argument.

USER STUDY

In this section, we present results from a preliminary study of undergraduate students at (anonymized university) who were assigned to participate in DebateCAFE. The resulting dataset contains responses from 94 students with 284 IBA responses, 170 new arguments on both sides and 1754 peer-to-peer ratings. Note that the population in this study is not a representative sample but this informal study provides insights into the performance of DebateCAFE and suggestions for future design choices. For analysis purpose, we scale the raw ratings to 0-1.

We first reveal the distribution of the IBA responses and the initial positions of the participants. Next we present the top arguments for both positions to evaluate their quality. After that, we quantify selective exposure and measure the subsequent biases. Finally we summarize the participant scores to evaluate the incentive mechanism.

The Initial Bias Assessment Questions

The first two questions provide a rough indication of the participant's initial attitude toward the issue and the third question captures how open the participant is to deliberation. Figure 3 panels (a), (b) and (c) show the histogram of each of the three IBAs. The response to IBA1 has a wider spread than IBA2 and IBA3 with a mean at 0.54, indicating people have varied preference in giving up privacy for security. IBA2 is skewed left suggesting that most participants value personal privacy and regard it as a constitutional right. For IBA3, responses are clustered at 0.7-1.0, reflecting an appreciation of arguments on both sides of an issue.

The relation between IBA1 and IBA2 is interesting. Figure 3(d) shows the scatter plot of IBA2 vs. IBA1. Observe that most points lie above the y=1-x line, demonstrating that participants who were not willing to give up privacy for security had a strong pursuit of personal privacy. Participants on the top left of the plot were likely to favor Apple's position, whereas participants on the bottom right were likely to favor the FBI. The participants' initial estimated positions were determined by the following metric:

If IBA2 - IBA1 $\geq$ 0.3, this participant is "pro-Apple";

If IBA1 - IBA2 $\geq$ 0.1, this participant is "pro-FBI";

These criteria were not symmetric because the ratings received for IBA1 and IBA2 were asymmetric and we wanted the two groups to have similar number of participants. This metric resulted in 27 "pro-Apple" participants, 23 "pro-FBI" participants and 44 "neutral" participants as shown in three colors in Figure 3 panel (d).

Arguments on Both Sides

Out of a total of 170 arguments, after ranking with the Wilson metric, 17 of the top 20 were "pro-Apple" arguments and 3 were "pro-FBI" arguments. This outcome may be a result of the demographic characteristics of the respondents, most of whom were university students. They were more likely to own an Apple product and may be more skeptical of government agencies. Here we present the top 3 arguments on both sides of the issue.

Pro-Apple (personal security) arguments:

1. Apple should not cooperate because it sets a dangerous precedent with regard to privacy and the FBI and other investigative groups. It also has the potential for issues with the same opening being exploited by hackers.
2. Apple should not cooperate because allowing the FBI access to these systems opens a door that cannot be closed again at will. Creating a cyber-security loophole allowing access for the FBI means that anyone with the technological knowledge can also exploit this access; a weakness in the fundamental security creates vulnerability. Our personal information would be vulnerable to sophisticated attacks by hackers and terrorists themselves.
3. Apple should not cooperate because it violates customer's privacy. If customers do not trust Apple with their data anymore then sales will drop and Apple will no longer be relevant in tech. One thing that the government is asking is to create a backdoor for a particular OS, which could lead to hackers exploiting a bug in subsequent OS's. This is a huge security red light because Apple does not and should not knowingly create a backdoor which could lead to major security concerns in the future.

Pro-FBI (national security) arguments:

1. Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.
2. Apple should cooperate because if there is a little compromise involved in guaranteeing the security of the nation, then as a resident and/or citizen of this nation, people should be willing to make that compromise.
3. Apple should cooperate because there is an increased security risk when they do not cooperate. If potential terrorist information is on the phones of those like the San Bernardino shooters, it will jeopardize the entire safety and security of the United States.

These arguments are rather well articulated with a clear position and concrete reasoning, demonstrating the capability of DebateCAFE in harvesting persuasive arguments for opposing stances. However, argument duplication is an issue: we observe many arguments conveying the same idea with slightly different wording. For example, the top 2 "pro-Apple" arguments both mention that complying with the FBI's request could lead to potential cyber-attack from hackers and the top 2 "pro-FBI" arguments point out the tradeoff between personal privacy and increasing national security from terrorist acts.

Clustering

Measuring Selective Exposure

As mentioned earlier, selective exposure is a major concern for online deliberation platforms, and DebateCAFE is able to quantify the extent of this behavior. Recall that DebateCAFE's peer-to-peer argument evaluation interface presents 8 arguments covering both sides of an issue. Each argument is represented by a sphere in the space with a tag of either "Apple" or "FBI" indicating the position of the argument. Participants are free to click on any sphere in the space when they first land on this page. This design enables us to observe which side of the argument a participant chooses to view first given his/her initial bias estimated by the IBAs. **20/27 pro-Apple participants first selected an argument for "Apple" and 11/23 pro-FBI participates first selected an argument for "FBI."** Participants were more inclined to first view an argument for "Apple" despite their initial position and a greater percentage of pro-Apple participants were more inclined to first view an argument that agreed with their position compared to pro-FBI participants. The difference was, however, not significant. Furthermore, **25/27 pro-Apple participants and 22/23 pro-FBI participants viewed at least one argument for the opposing position**. This high percentage suggests that indirectly exposing arguments on both sides via an uncertainty-sampling interface can prompt participants to proactively view/rate adversarial arguments.

Articulation Bias

Here we would like to compare the quality of participants' arguments for and against their initial position (as estimated from the results of their IBAs). Peer-ratings of argument persuasiveness reveal that **among the pro-Apple participants who provided rated arguments, only 3/20 had a higher rated argument for "FBI." However, among the pro-FBI participants who provided rated arguments, 8/21 had a higher rated argument for "FBI."** Some participants did not provide arguments and some arguments were not rated. Results indicate that pro-Apple participants were less likely to articulate persuasive arguments for the opposing view, while pro-FBI participants had little trouble crafting persuasive arguments for the opposing view, but the difference is not significant.

Peer Rating Bias

We received 1754 valid peer ratings, among which 851 rated "pro-FBI" arguments and 903 rated "pro-Apple" arguments. By adopting the Welch two-sample t-test, with a t-value of 4.289 and p-value <0.01, we conclude that the **"pro-Apple" arguments were, on average, receiving significantly higher ratings than "pro-FBI" arguments.** We conjecture this difference came from the inherent bias of participants, who tended to rate more highly those arguments that align with their position.

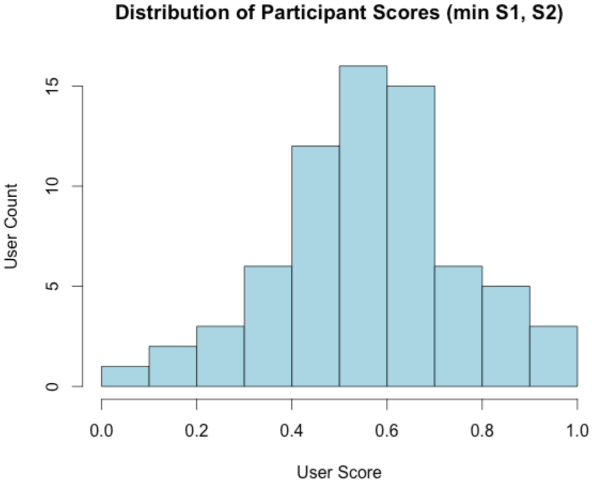


Figure 4: Histogram of scaled scores of all participants.

Furthermore, we classify the arguments as either "Consistent" or "Adversarial" with respect to initial bias. From participants in the pro-Apple and the pro-FBI groups, "Consistent" arguments received an average rating of 0.523 with a Standard Deviation (SD) of 0.25 and a Standard Error (SE) of 0.0118, while "Adversarial" arguments received an average rating of 0.485 with a SD of 0.25 and a SE of 0.0119. Although the absolute difference was under 5%, with a p-value of 0.022, the Welch two-sample t-test suggests that the existed a true difference in means, suggesting a **consistency between initial bias and ratings bias.** However, a controlled experiment is warranted to explore this.

Participant Score

DebateCAFE defines participant score as the minimum of the two argument scores: $\min(s_1, s_2)$, i.e., based on the weaker of the two arguments entered, to reflect a participant's ability to articulate adversarial arguments. We scaled all the scores to 0 to 1 for more intuitive comparison. From Figure 4, we observe that participant scores followed a normal distribution, where few participants were extremely capable or incapable of articulating adversarial arguments and most participants were mediocre. Among the top-scoring participants, we confirm they were able to provide strong adversarial arguments. For example, the two arguments from a top-scoring participant were:

(Pro-FBI argument) Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.

(Pro-Apple argument) Apple should not cooperate because if this gets scaled to all cases, it could seriously diminish the customers' privacy.

Note the first argument was among the top 3 pro-FBI arguments and the second argument focused on the consequence of scaling such action to all cases.

Among the low-scoring participants, the Wilson scores of their two arguments were highly lopsided. The participant with the lowest score only provided a pro-Apple argument, leaving the pro-FBI argument as "Apple should cooperate because." Another participant with a low score gave a "pro-FBI" argument as "Apple should cooperate because public security?" while providing a well-articulated "pro-Apple" argument: "Apple should not cooperate because it has paid such amount of advertisement and technology to increasing security, so it should not yield all to FBI".

DISCUSSION

We present a novel deliberation platform, DebateCAFE, designed to encourage users to review and articulate adversarial arguments on contentious issues. DebateCAFE implements collaborative filtering to identify persuasive arguments and to balance exposure to polarized viewpoints using uncertainty sampling. By introducing a scoring system to users, DebateCAFE incentivizes them to articulate persuasive arguments on both sides of the issue.

We describe an application of DebateCAFE to the Apple (personal privacy) vs. FBI (national security) issue and report results. By presenting an equal number of arguments on both positions using uncertainty sampling, our platform is successful in motivating participants to view and rate arguments on opposing positions, mitigating selective exposure and achieving a more balanced discussion around the issue. Even though a small sample size did not permit us to show that DebateCAFE helped participants change their minds on this particular issue, results suggest that DebateCAFE can reduce selective exposure bias without relying on high-cost human moderation.

FUTURE WORK

De-duplication

During the deployment, we found that argument duplication was a significant problem. Many participants provided similar arguments with slightly different wording. For example, among the "pro-Apple" arguments, many identified the potential that hackers could gain access to all Apple devices. Duplicate arguments should be consolidated to optimize the effectiveness and efficiency of participants' peer-ratings. One possible solution is to introduce a moderator, who reads and consolidates arguments. However, when the platform scales, it would be infeasible for the moderator to view every single argument. An alternative approach is to identify potentially similar arguments using Natural Language Processing techniques, and then enlist participants (in an interface similar to that used in the Argument Articulation Phase) to de-duplicate or synthesize those arguments.

Enhancing Value of Persuasiveness

As we mentioned before, the value for being persuasive varies for different individuals. To further improve the effectiveness of the persuasive scoring mechanism, we would like to tie this score back to the design of the platform. One approach is to introduce a positive relationship between argument exposure and user persuasiveness, where arguments provided by more persuasive users have a higher probability of being shown. Alternatively, we may completely hide those arguments provided by users with low persuasive score. Another approach is to add a public leader board with a list of top users ranked by their persuasive score.

Measuring Changes in Opinion

When participants rate persuasiveness of arguments, they might give high ratings to arguments that are logically coherent but are peripheral to the issue or of marginal importance. To better measure "persuasiveness" as opposed to "logical validity," we will experiment with a slider indicating the participant's position on the issue. It will be available throughout the discussion phase, allowing the participant to record her/his opinion change after viewing each argument. The persuasiveness of each argument will then be captured by the total opinion changes of other participants.

ACKNOWLEDGEMENT

This work is supported in part by the Blum Center for Developing Economies and the Development Impact Lab (USAID Cooperative Agreement AID-OAA-A-12- 00011) as part of the USAID Higher Education Solutions Network, (anonymized university)'s Algorithms, Machines, and People (AMP) Lab, and the Connected Communities Initiative at the Center for Information Technology Research in the Interest of Society (CITRIS) and the Banatao Institute at (anonymized university). This work is also supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (no. IIID-2015-07). Thanks to all students who participated and colleagues and reviewers who provided feedback on earlier drafts.

REFERENCES

1. Steffen Albrecht. 2006. Whose voice is heard in online deliberation?: A study of participation and representation in political debates on the internet. *Information, Community and Society*, 9 (1). 62-82.
2. Matt Apuzzo, Joseph Goldstein and Eric Lichtblau. 2016. Apple's Line in the Sand Was Over a Year in the Making, The New York Times.
3. Peter Baldwin and David Price. Debate-Graph Details.
4. Lucian Bebchuk and Jesse Fried. 2004. Pay without performance, Cambridge, MA: Harvard University Press.
5. Jeffrey P Bigham, Michael S Bernstein and Eytan Adar. 2015. Human-computer interaction and collective intelligence. *Handbook of Collective Intelligence*, 57.
6. Andrea Bunt, Matthew Lount and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, ACM, 169-178.
7. Yan Chen, F Maxwell Harper, Joseph Konstan and Sherry Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *The American economic review*, 100 (4). 1358-1398.
8. Lydia B Chilton, Juho Kim, Paul André, Felicia Cordeiro, James A Landay, Daniel S Weld, Steven P Dow, Robert C Miller and Haoqi Zhang. 2014. Frenzy:

collaborative data organization for creating conference sessions. in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM, 1255-1264.

9. Todd Davies and Seeta Peña Gangadharan. 2009. Online deliberation: Design, research, and practice.
10. Peter Denning, Jim Horning, David Parnas and Lauren Weinstein. 2005. Wikipedia risks. *Communications of the ACM*, 48 (12). 152-152.
11. Kathleen M Eisenhardt. 1989. Agency theory: An assessment and review. *Academy of management review*, 14 (1). 57-74.
12. Siamak Faridani, Ephrat Bitton, Kimiko Ryokai and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1175-1184.
13. Mingkun Gao, Hyo Jin Do and Wai-Tat Fu. 2017. An Intelligent Interface for Organizing Online Opinions on Controversial Topics. in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, ACM, 119-123.
14. Eric Gilbert. 2013. Widespread underprovision on Reddit. in *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 803-808.
15. David Goldberg, David Nichols, Brian M Oki and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35 (12). 61-70.
16. Eduardo Graells-Garrido, Mounia Lalmas and Ricardo Baeza-Yates. 2016. Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles. in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ACM, 228-240.
17. Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM, 169-180.
18. Michael C Jensen and William H Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3 (4). 305-360.
19. Arjun Kharpal. 2016. Apple vs FBI: All you need to know, CNBC.
20. Mark Klein. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.
21. Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, 36 (3). 426-448.
22. Tetsuro Kobayashi and Ken'ichi Ikeda. 2009. Selective exposure in political web browsing: Empirical verification of 'cyber-balkanization'in Japan and the USA. *Information, Communication & Society*, 12 (6). 929-953.
23. Kenneth L Kraemer and John Leslie King. 1988. Computer-based systems for cooperative work and group decision making. *ACM Computing Surveys (CSUR)*, 20 (2). 115-146.
24. Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning and Lance Bennett. 2012. Supporting reflective public thought with considerit. in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, 265-274.
25. Sanjay Krishnan, Yuko Okubo, Kanji Uchino and Ken Goldberg. 2013. Using a social media platform to explore how social media can enhance primary and secondary learning. in *Learning International Networks Consortium (LINC) 2013 Conference*.
26. Sanjay Krishnan, Jay Patel, Michael J Franklin and Ken Goldberg. 2014. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. in *Proceedings of the 8th ACM Conference on Recommender systems*, ACM, 137-144.
27. Jean-Jacques Laffont and David Martimort. 2009. *The theory of incentives: the principal-agent model*. Princeton university press.
28. Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now?: mitigating the echo chamber effect by source position indicators. in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, 184-196.
29. Q Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. in *Proceedings of the 32nd annual ACM conference on human factors in computing systems*, ACM, 2745-2754.
30. Greg Linden, Brent Smith and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7 (1). 76-80.
31. Mark EJ Newman. 2001. Clustering and preferential attachment in growing networks. *Physical review E*, 64 (2). 025102.
32. Brandie Nonnecke, Sanjay Krishnan, Jay Patel, Mo Zhou, Laura Byaruhanga, Dorothy Masinde, Maria Elena Meneses, Alejandro Martin del Campo, Camille Crittenden and Ken Goldberg. 2015. DevCAFE 1.0: A participatory platform for assessing development initiatives in the field. in *Global Humanitarian Technology Conference (GHTC), 2015 IEEE*, IEEE, 437-444.
33. Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.
34. Christian Pentzold. 2011. Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their 'community'? *New Media & Society*, 13 (5). 704-721.
35. Vincent Price and Joseph N Cappella. 2002. Online deliberation and its influence: The electronic dialogue project in campaign 2000. *IT & Society*, 1 (1). 303-329.
36. Eric Rasmusen and Basil Blackwell. 1994. Games and information. *Cambridge, MA*, 15.
37. Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe and Kristy Milland. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 1621-1630.
38. Matthew J Salganik, Peter Sheridan Dodds and Duncan J Watts. 2006. Experimental study of inequality and

unpredictability in an artificial cultural market. *science*, 311 (5762). 854-856.

39. Cristina Sarasua, Elena Simperl and Natalya F Noy. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. in *International Semantic Web Conference*, Springer, 525-541.

40. Douglas Schuler. 2009. Online civic deliberation with e-Liberate. *Online deliberation: Design, research, and practice*. 293-302.

41. David O Sears and Jonathan L Freedman. 1967. Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31 (2). 194-213.

42. Joanna E Siegel, Jessica Waddell Heeringa and Kristin L Carman. 2013. Public deliberation in decisions about health research. *Virtual Mentor*, 15 (1). 51.

43. Stephanie Solomon and Julia Abelson. 2012. Why and when should we use public deliberation? *Hastings Center Report*, 42 (2). 17-20.

44. Abhay Sukumaran, Stephanie Vezich, Melanie McHugh and Clifford Nass. 2011. Normative influences on thoughtful online participation. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3401-3410.

45. W Ben Towne and James D Herbsleb. 2012. Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, 9 (1). 97-115.

46. Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22 (158). 209-212.

47. Svante Wold, Kim Esbensen and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2 (1-3). 37-52.

48. Xiwang Yang, Yang Guo, Yong Liu and Harald Steck. 2014. A survey of collaborative filtering based social recommender systems. *Computer Communications*, 41. 1-10.

49. Mo Zhou, Alison Cliff, Sanjay Krishnan, Brandie Nonnecke, Camille Crittenden, Kanji Uchino and Ken Goldberg. 2015. M-CAFE 1.0: Motivating and Prioritizing Ongoing Student Feedback During MOOCs and Large on-Campus Courses using Collaborative Filtering. in *Proceedings of the 16th Annual Conference on Information Technology Education*, ACM, 153-158.