Opinion Space: A Scalable Tool for Browsing Online Comments

Siamak Faridani[1], Ephrat Bitton[1], Kimiko Ryokai[3], Ken Goldberg[1,2,3]

[1]IEOR Department
UC Berkeley
Berkeley CA 94720
{faridani, ebitton}@berkeley.edu

[2]EECS Department
UC Berkeley
Berkeley CA 94720
goldberg@berkeley.edu

[3]School of Information
UC Berkeley
Berkeley CA 94720
kimiko@ischool.berkeley.edu

ABSTRACT

Internet users are increasingly inclined to contribute comments to online news articles, videos, product reviews, and blogs. The most common interface for comments is a list, sorted by time of entry or by binary ratings. It is widely recognized that such lists do not scale well and can lead to "cyberpolarization," which serves to reinforce extreme opinions. We present Opinion Space: a new online interface incorporating ideas from deliberative polling, dimensionality reduction, and collaborative filtering that allows participants to visualize and navigate through a diversity of comments. This self-organizing system automatically highlights the comments found most insightful by users from a range of perspectives. We report results of a controlled user study. When Opinion Space was compared with a chronological List interface, participants read a similar diversity of comments. However, they were significantly more engaged with the system, and they had significantly higher agreement with and respect for the comments they read.

Author Keywords

Comment Browsing, Opinion Mining, Deliberative Polling, Collaborative Filtering, Dimensionality Reduction, Perceptual Maps, Opinion Visualization, Very Large Scale Conversations

ACM Classification Keywords

H.5.2 User Interfaces, H.5.3 Group and Organization Interfaces

General Terms

Algorithms, Design, Experimentation

INTRODUCTION

"Opinion is the medium between ignorance and knowledge." - Plato

A central aspect of "participatory culture" is that users of online sites for news, blogs, videos, and commerce increasingly provide feedback in the form of textual comments. While participatory culture thrives on the sharing of diverse opinions among large populations over the network, there are several problems with existing systems. First, thoughtful moderates are often shouted down by extremists. Online discussions, conducted through threaded lists of comments, often end in "flame wars" predicated on binary characterizations. Second, the amount of data can be overwhelming. News stories and blog posts often generate hundreds or thousands of comments. As the number of comments grows, presenting them in a chronological list is simply not a scalable interface for browsing and skimming. Third, many websites tend to attract people with like-minded viewpoints, which can reinforce biases and produce "cyberpolarization" [22].

Discourse Architecture is the study and design of technologies that facilitate very large-scale conversations. This area of study is related to Computer-Supported Cooperative Work (CSCW), Computer-Human Interaction (CHI), and Computer-Mediated Communication (CMC). Discourse Architecture studies of online discussion forums have recognized the limitations of linear comment lists [19, 23].

Opinion Space (accessible at http://opinion.berkeley.edu) is a new online tool designed to collect and visualize user opinions on topics ranging from politics to parenting, from art to zoology. With Opinion Space, we aim to address the above problems by incorporating ideas from deliberative polling, dimensionality reduction, and collaborative filtering. Opinion Space solicits opinions to a set of controversial statements as scalar values on a continuous scale (from strongly disagree to strongly agree) and applies dimensionality reduction to project the data onto a two-dimensional plane for visualization and navigation, effectively placing all participants onto one level playing field. Points far apart correspond to participants with very different opinions, and participants with similar opinions are proximal. One of our goals is to move beyond one-dimensional characterizations of opinion: the arrangement

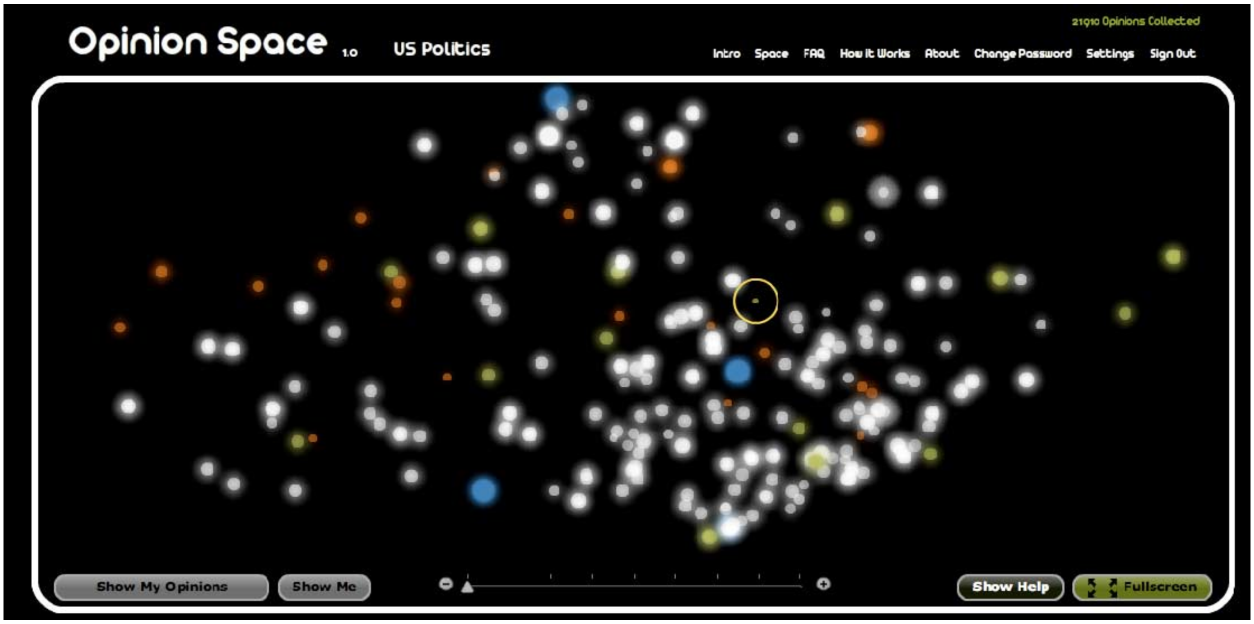

Figure 1. A screenshot of the Opinion Space 1.0 interactive map. Each point corresponds to a user and comment. The point with the halo indicates the position of the active user; green points correspond to comments rated positively by the active user, and red points correspond to comments rated negatively. Larger and brighter points are associated with the comments that are rated more positively by the user community.

of points is statistically optimized to convey the underlying distribution of opinions and does not correspond to conventional left/liberal and right/conservative polarities. Participants are also asked to contribute a textual comment in response to a discussion topic; each comment is associated with the position of the contributing user in the visualization space. We designed Opinion Space to be a self-organizing system that rewards participants who consider the opinions of those with whom they might normally disagree.

The first version of Opinion Space (v1.0) was released to the general public on March 28, 2009. In the first few months, it attracted 21,563 unique visitors of which 4,721 registered with their email address with the purpose of saving their settings. In this "in the wild" experiment, each registered user rated on average 14.2 comments.

The positive response to Opinion Space motivated us to conduct a controlled user study to quantify and compare Opinion Space with other interfaces in terms of user engagement and the ability to find valuable comments. In this paper we present the background research, the user interaction design and implementation of Opinion Space, and the design and results of the user study.

RELATED WORK

There is a large body of literature on the challenges of online discussion.

Deliberative Polling and Opinion Mining

Dahlgren [7] argues that one of the dangers of online deliberations is fragmentation of the participants. While Berinsky [1] lauds public opinion polling as one of the most inclusive means for participating in political discussions, he is critical of its inherent bias.

First proposed by Fishkin in 1991 [9], deliberative polling is an alternative to traditional polling techniques where participants are first polled on a set of issues, allowed to deliberate for a period of time, and then polled once more. The outcome is often a better understanding of how public opinion would change if people were more informed on the issues. Opinion Space can be thought of as an online, asynchronous version of deliberative polling, where users can inform each other and adjust their opinions over time.

Pang and Lee [17] survey several techniques for gathering and understanding political and consumer opinions. Specifically, they review the literature on the problems of identifying opinionated material in a document, determining the underlying sentiments of the material, and summarizing the information in an effective way.

Visualizing Social Networks

Opinion Space defines a metric relationship between users based on similarity of opinion, which lends itself well towards forming a geometrically meaningful visualization of the users in a two-dimensional plane. An underlying network structure emerges in this space as users interact by rating each other's comments.

The structure of social networks is an active area of research [4]. Freeman [8] provides background on visualization in social network analysis, from hand-drawn to computer-generated. Viegas and Donath [25] explore two visualizations based on email patterns: a standard graph-based visualization and a visualization that depicts temporal rhythms. They found that the latter complements and enhances the former, suggesting that going beyond visualization of relationships in the graph, which is what we aim to do with Opinion Space, is a more effective way to explore and analyze interactions in social networks.

There are several systems available that were designed to aid in the analysis of social networks by providing effective visualization and navigation capabilities. Morningside Analytics (http://morningside-analytics.com/) is a company that develops powerful tools for mapping and visualizing emerging trends in online communities using textual analysis. Sack presents the *Conversation Map* interface that analyzes messages using a set of computational linguistics and sociology techniques to generate a graphical display of links between messages based on textual content [20]. Other visualization interfaces include SocialAction, which, like Opinion Space, allows for the visualization of several social network analysis measures [18]. Vizster is a system for visual search and structure analysis [13]. Like Opinion Space, Vizster uses proximity to highlight similarity. However, Vizster is based on binary connectivity models and does not represent gradations of opinion.

Increasing Participation in Online Communities

Bishop [2] presents a theoretical framework for understanding what factors encourage visitors to participate in online discussion forums, and Brandtzaeg and Heim [6] describe a user study of participation in several popular Norwegian online forums. Ludford et al. [15] found that online participation increases when the groups formed are diverse and when users are told how unique they are to the group. Opinion Space builds on both Bishop's framework and Ludford's findings for more positive participation online by visualizing a spectrum of opinions that is much broader than binary differences.

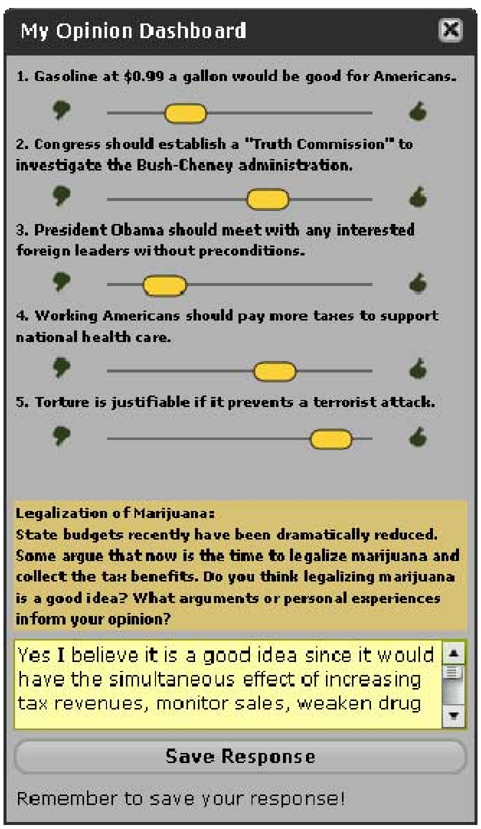

Figure 2. Users indicate their opinions on five profile propositions using horizontal sliders and type a textual comment related to the current discussion question.

Identifying Insightful Comments

One of the key challenges Opinion Space faces is to identify the most insightful comments based on ratings collected from the users. Thus it is related to collaborative filtering, a technique used by sites such as Netflix, Amazon, Digg, and our Jester joke recommender system to make recommendations by combining user ratings [12].

Opinion Space also draws on several existing applications designed to give political voting advice. In 2008, the Washington Post released Poligraph, an online application that plotted the US presidential candidates on a two-dimensional graph with respect to their stances on several healthcare reform issues. After responding to a series of questions, users' stances were plotted in comparison. EU Profiler (http://www.euprofiler.eu) is an online voting application designed to help users better understand the political landscape of member states of the European Union and to determine where they stand within it. Unlike Opinion Space, both Poligraph and EU Profiler are not collaborative and they do not model the distance between users.

DESIGN OF OPINION SPACE 1.0

In this section we describe the design of Opinion Space 1.0 by stepping through the experience of a new user. We then describe and motivate the methods used to create the Opinion Space visualization.

User Activities

Entering Opinions

As illustrated in Figure 2, a new user is presented with five "opinion profile" propositions and asked to rate them on a continuous scale between "strongly disagree" and "strongly agree." All ratings are entered via a horizontal "slider" that is operated like a scroll bar. The first version of Opinion Space (1.0) focused on issues related to US domestic politics. As shown in Figure 2, the initial propositions addressed issues such as the price of gasoline, and the discussion question prompted users to contribute textual comments regarding the benefits and consequences of legalizing marijuana. The propositions are designed to elicit a diversity of viewpoints (i.e., have high variance in responses). The user is also asked to enter a textual comment on the current discussion topic. Users are free to change the ratings in their opinion profiles and edit their comments at any time.

Browsing and Rating Comments

The user is then presented with the Opinion Space "map," a projection of a sample of the users onto a two-dimensional

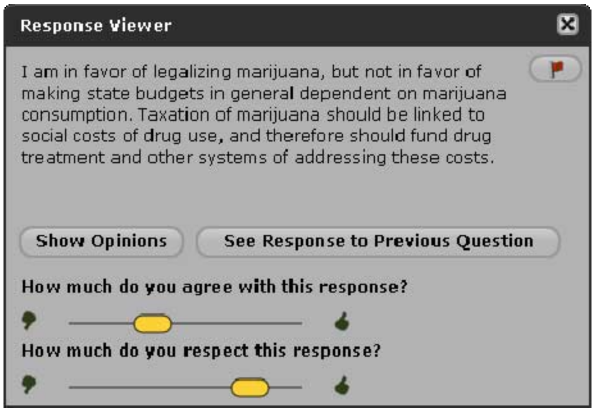

Figure 3. Users read each comment and rate it based on how much they agree with the comment and how much they respect it.

plane where each user is represented by a point based on the 5-dimensional response to the profile opinions (Figure 1). A yellow point surrounded by a halo indicates the location of the active user. Other users are initially displayed as white points until they are rated by the active user. Points far apart correspond to participants with very different opinions, and participants with similar opinions are near each other in the space.

Users can view and rate responses by clicking on other points in the Opinion Space map. When a point is selected, a window (Figure 3) will appear displaying the associated comment. Directly below the comment text, the user is prompted to indicate the degree of his or her agreement with and respect for the comment by using two slider bars. The size and brightness of each point is determined by a weighted average of the ratings that other users have assigned the corresponding comment and the distance in Euclidean space between those users and the commenter. Larger and brighter points correspond to the comments that are more agreeable to a diversity of users rather than those sharing similar beliefs; the specifics of our model for scaling ratings in this way is described in [3].

Dimensionality Reduction

Mathematically, the map is a projection of the five-dimensional opinion profiles onto two dimensions using a technique known as principal component analysis (PCA). Under assumptions of independence and linearity, PCA allows us to reduce the dimension of the opinion profiles

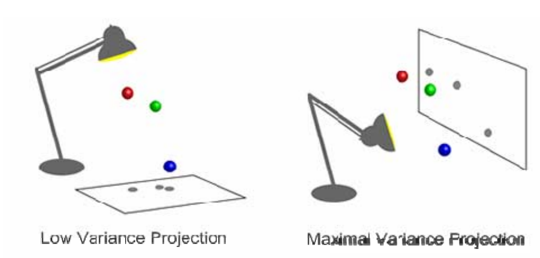

Figure 4. Dimensionality reduction from 3D to 2D.

while maximizing the variation in distance relationships between users [14]. Figure 4 is an illustrative example of the challenges of dimension reduction from 3D to 2D, using a light and shadows as a metaphor for projection; if done incorrectly, as shown in the low variance projection, distance relationships in the 2D projection can be the reverse of what is true in 3 dimensions.

Reducing a dataset to two dimensions via PCA can be summarized by finding the two largest eigenvectors (α_1, α_2) of the covariance matrix Σ of the data. These two eigenvectors account for the most variation of the data, and are referred to as the first two principal components. Given an opinion profile x, its corresponding coordinate in the Opinion Space map is given by the dot product of x and the eigenvector: $(x{\cdot}\alpha_1, x{\cdot}\alpha_2)$.

We chose to use PCA to build the Opinion Space map because it finds the projection that minimizes squared error, and the position of a new user can be computed in constant time. However, many other dimensionality-reduction techniques are known in the literature [10]; these include factor analysis, multi-dimensional scaling (MDS), singular value decomposition, projection pursuit, independent component analysis, and t-distributed stochastic neighbor embedding (t-SNE) [10, 24]. While there are many merits to these techniques, they are not as scalable or efficient as PCA. Some, such as t-SNE, use PCA as a pre-processing step to make larger problems computationally manageable.

USER STUDY

We created three interfaces, List, Grid, and Space (the last most similar to Opinion Space 1.0), and populated each with a set of 200 randomly selected user comments from the "in the wild" experiment. We presented each of the interfaces in random order to 12 study participants in a within-subject study using the Space interface as the experimental condition and the List and Grid interfaces as two control conditions, and we recorded data as the users read and rated the comments of others.

In the following subsections, we describe each of the three interfaces in greater detail, the hypotheses we formed regarding Opinion Space 1.0, and the protocol we followed for conducting the user study.

Three Interfaces Compared in Study

List Interface

The List interface (shown in Figure 5) is based on standard comment lists found on blogs and other websites. In the List interface, 200 comments are presented in a chronological linear list. We record the amount of time participants spend on every comment they view ("dwell time") as well as the *agree* and *respect* ratings they give to each comment. To more accurately measure the time users spend reading a comment, neighboring comments are blurred and then instantly de-blurred as the user scrolls up or down the list.

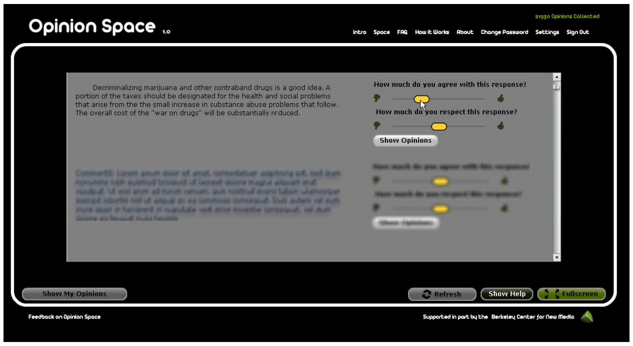

Figure 5. List interface. comments are presented in a list in chronological order

Figure 6 Grid interface. All of the comments are presented in a graphical grid in the chronological order

Grid Interface

The Grid interface (shown in Figure 6) is designed to be a control for studying the effect of visualizing the points based on the spread of opinion profile data. The Grid interface is a graphical display similar to Opinion Space 1.0, the primary difference being the positioning of the points. Here, points are ordered on a uniform rectangular grid according to time of entry; the location of a point is only a function of the time it was entered and is independent of the corresponding user's opinion profile. The size and brightness of the points varies with user ratings, as in the Space Interface. Study participants were asked to click on points in any order they wished and to rate the comments.

Space Interface

The Space interface is the experimental condition and is nearly identical to Opinion Space 1.0. We turned off cosmetic features such as the "twinkling" of points to avoid any bias they might introduce by unintentionally influencing which points users choose to click.

Hypotheses

We considered five hypotheses. Based on our design goals for Opinion Space 1.0, we expected that:

Hypothesis 1 (H1): Opinion Space will be significantly more engaging than List or Grid in terms of average dwell time [H1a] and in terms of user ranking of overall preference [H1b].

Since Opinion Space combines user ratings with metric information about relative opinion positions (See [3] for more details.) we expected that:

Hypothesis 2 (H2): Users will report Opinion Space as more conducive to finding "useful" comments than List or Grid interfaces.

An important goal for Opinion Space was to expose users to a wider range of insightful opinions rather than the majority view or the most recently posted comments. We measure the *diversity* of a comment encountered by user i as the Euclidean distance between user i's opinion profile and that of the user who wrote the comment. We define the diversity of a set of comments encountered by user i as the average pairwise Euclidean distance between i and each commenter in the set. Hence, our third hypothesis is that

Hypothesis 3 (H3): Users of Opinion Space will read a significantly more diverse set of comments than with the List or Grid interfaces.

Since Opinion Space is designed to highlight the most insightful comments by increasing the size and brightness of the corresponding points in the map, we expect that users will find and read more comments they agree with when using the Space interface.

Hypothesis 4 (H4): Opinion Space users will report significantly great agreement with the comments of others than they do when using the List or Grid interfaces.

Finally, motivated by the notion that it is easier to respect the opinion of an individual given more contextual information (such as the political views of that person), we expect that

Hypothesis 5 (H5): Opinion Space users will report significantly greater respect for the comments of others than they do when using the List or Grid interfaces.

Method

To test our hypotheses, we designed a within-subject study using the Space interface as the experimental condition and the List and Grid interfaces as two control conditions. Each participant interacted with all three interfaces, and the interfaces were presented in random order so as to reduce the potential for bias.

Question	Mean		Variance
Age	19.9	±	0.9
How tech savvy are you?	5.9/10	±	4.9
How familiar are you with the current political issues?	6.0/10	±	3.0

Table 1. Characteristics of 12 Study Participants.

	List	Grid	Space
Average number of Comments Rated	23.5 ± 11.2	20.9 ± 9.9	21.1 ± 9.0
Average Dwell Time per Comment (sec)	516.4 ± 242.5	458.4 ± 180.4	582.9 ± 187.1
Average "Agree with" Rating	0.443 ±0. 266	0.515 ± 0.278	0.567 ± 0.269
Average "Respect for" Rating	0.396 ±0. 294	0.479 ± 0.300	0.510 ± 0.284

Table 2. Average Data for 12 study participants (mean ± standard deviation)

	List	Grid	Space
I found this version of the system enjoyable to use.	2.2 ± 1.3	3.3 ± 1.2	4.8 ± 0.4
I learned something interesting while using this version.	2.9 ± 0.9	3.6 ± 0.9	4.2 ± 0.7
This version is conducive towards finding useful comments.	2.0 ± 1.2	3.3 ±0.8	4.2 ± 0.7

Table 3. Average response data from short questionnaires asking users to indicate how enjoyable, interesting, and useful they found each interface by providing an integer-valued rating from 1 (not at all) to 5 (very much).

User Study Participants

12 participants were selected from a pool of 36 volunteers who responded to our ads posted across the UC Berkeley campus and Facebook. All of the volunteers in that pool completed an online pre-screening survey to ensure that they were not already familiar with Opinion Space 1.0 and that they had a relatively good understanding of current political issues in the US. Each participant was offered a $10 gift certificate to Amazon.com for successfully completing the experiment.

We had two female and ten male volunteers participate in the study. Three participants identified themselves as Republican (25%), five as Democrats (42%) and 4 as Independents (33%). Additional information about the participants is provided in Table 1.

Experiment Protocol

Each individual experiment took approximately one hour to complete. Sessions began by having participants use the proposition sliders to enter their own opinion profiles and by having them enter a textual comment on the current discussion question regarding the legalization of marijuana. They were then asked to explore each of the three interfaces, which were presented to them in random order. Participants were free to switch to the next interface whenever they wanted so long as they had rated at least 10 comments; we wanted to ensure that participants had at least a minimal amount of experience interacting with each interface. If a participant did not ask to switch to the next interface after 15 minutes, the system did so automatically.

After using each interface and before moving on to the next, participants were given a short questionnaire that asked them to indicate on an integer scale of 1 (not at all) to 5 (very) how enjoyable, interesting, and useful they found the interface. Participants were encouraged to explore each interface freely by reading and rating comments in any order they wished. We automatically recorded user dwell time for each comment. Participants were asked to read comments carefully and rate them individually based on how much they agree with the comment and how much they respect it (Figure 3). Upon completion of the experiment, participants were given an exit survey that asked them to rank the three interfaces on a series of 7 qualities.

SUMMARY OF RESULTS

In this section we describe the results of our study as determined both objectively with numerical, observational data and subjectively through questionnaires completed by the participants. Table 2 shows the mean and standard deviation of the number of comments rated by the participants in each of the three interfaces. The third and fourth rows show the average participant rating of each comment on a continuous scale between 0 and 1, in terms of the *agree* and *respect* measures, respectively.

Table 3 summarizes the mean and standard deviation of participant responses to the short questionnaire asking users how enjoyable, interesting, and useful they found each interface by providing an integer value from 1 (not at all) to 5 (very much). Table 4 summarizes data from the exit survey that asked participants to rank the interfaces after trying all three.

Carry-over effect of User Fatigue

Interfaces were presented in random order for each participant. To check for the presence of carry-over effects between interfaces due to user fatigue, we recorded the total time users spent with each interface as a measure of engagement. We conducted a two-way ANOVA analysis on the distributions of the time users spent with the first, second, and third interfaces presented to them. Our analysis yielded a p-value of $0.534 >> 0.05$, which suggests that user fatigue did not cause significant carry-over effects.

Evaluation of Hypotheses

To analyze study data, we used Analysis of Variance (ANOVA), ANOVA on Ranks, Student t-tests, Friedman's test, Welch's test, and the Wilcoxon signed-rank test for significance, as well as Bartletts test for homogeneity of variance. ANOVA generalizes the Student t-test for measuring the statistical significance of the differences between data sets by analyzing their relative means and variances. Given *n* data sets, these tests produce a "p-value" that estimates the probability that the outcome is by chance, ie, that the sets were sampled from the same distribution; known as the *null hypothesis*. Lower p-values correspond to greater significance of the data.

Performing ANOVA reduces the chances of encoutering type I errors that may occur in executing multiple t-test hypothesis testing [16]. Similar to the Student t-test, ANOVA assumes that the observations are normally distributed and that the variances are equal. Before performing ANOVA, we use Bartlett's test to make sure that the homogeneity of variances (homoscedasticity) property holds. If the p-value for this test is high, we can perform an ANOVA analysis on the dataset. For analyzing ranked data (as with hypotheses H1a, H2) we use Friedman's test, which is an extension of ANOVA for nonparametric data [16].

Hypothesis 1 (H1): Opinion Space will be more significantly engaging than List or Grid in terms of average dwell time [H1a] and based on user ranking of all three interfaces in terms of overall preference [H1b].

We recorded dwell times for the 959 comments viewed by the participants while working with the three interfaces (users did not rate all comments they read). There are 329, 285, and 345 dwell times for the List, Grid and Space interfaces respectively. Average dwell times for these interfaces are reported in Table 2.

Bartlett's test rejected the assumption of homogeneity of variances for the dwell times, and so we performed a two-way, within-subject *ANOVA on Ranks* as suggested by [5]. For our analysis, the within-subject factor is the type of interface. The resulting p-value (1.098×10^{-14}), is significantly less than 0.05 suggesting that the type of interfaces impacted user dwell times. We used Welch's t-test to measure the extent of this impact, which is a generalization of the Student's t-test for cases where the variances are not equal [26]. Pairwise analysis using Welch's test shows that the dwell times in Grid and Space interfaces are significantly longer than the List interface (p-values for Grid-List is 2.2×10^{-16} and is 5.387×10^{-10} for List-Space , both << 0.05). However, we did not find a significant difference in the dwell times between the Grid and Space interfaces (p-value=0.1126 > 0.05).

We also performed Freidman's test on user responses to the question: "In which version do you expect to spend more time reading comments?" (Table 4). Freidman's test on this data yields a p-value of 0.0000984 << 0.05. We used Wilcoxon's signed-rank test as a pairwise post-test for nonparametric distributions. The test showed statistical significance between the user reported ranks for each pair of interfaces (p-values are: 0.02332 for Grid-List, 0.02351 for Grid-Space and 0.002608 for Space-List), which supports H1a.

The self-reported, subjective data suggests that users are significantly likely to spend more time reading comments on the Space interface, but the observed (objective) data does not show a significant difference between the Space and Grid interfaces.

To assess hypothesis H1b, we consider the data collected from the exit survey question that asked participants to rank the three interfaces by preference. Almost all (92%) of participants reported that they prefer Opinion Space to the List and Grid interfaces (*H1b*), as shown in Table 4. Friedman's ANOVA analysis on this data produces a p-value = 0.000486512 << 0.05, and Wilcoxon's signed-rank post-test shows statistical significance between each pair of user interfaces with p-values < 0.05. The results of this analysis mildly support hypothesis H1b. (for List-Space p-value= 0.01188, for Grid-Space p-value= 0.03884 and for Grid-List, p-value = 0.0209).

Hypothesis 2 (H2): Users will report Opinion Space is more conducive to finding "useful" comments than List or Grid interfaces.

In the questionnaires following the use of each interface, participants subjectively reported Opinion Space to be more conducive to finding useful comments than the List and Grid interfaces (Table 3). Conducting Freidman's test on this ranked data yields a p-value = 0.00361 << 0.05. Wilcoxon's post-test suggests that statistical significance holds for all pairs of interfaces (p-values for the follow up tests are: 0.003583 for Grid-PCA, 0.01868 for List-PCA and 0.03667 for Grid-List), in support of H2.

Hypothesis 3 (H3): Users of Opinion Space will read a significantly more diverse set of comments than with the List or Grid interfaces.

As noted earlier, we define the average diversity of a set of comments rated by user *i* as the average Euclidean distance between user *i* and the authors of those comments. In the 5D opinion profile vector space, the maximum distance between any two participants is 2.23 units. The average diversity for the 959 comments read by the 12 participants was 0.960, 0.924, and 0.992 for the Space, List, and Grid interfaces respectively. The data passes Bartlett's test for homogeneity of variances with a p-value of 0.1628 > 0.05, and ANOVA yields a p-value of 0.7848 >> 0.05. This suggests that there is no statistically significant difference between the diversity of comments read in each interface; hence, the data does not support H3.

Interestingly, participants (subjectively) *perceived* greater comment diversity in Opinion Space. In the exit Survey, 50% of participants reported Opinion Space allowed them to see more diverse comments; while only 16 % chose List and 33% chose Grid, as indicated by Question 6 in Table 4.

Hypothesis 4 (H4): Opinion Space users will report significantly great agreement with the comments of others than they do when using the List or Grid interfaces.

Question	List	Grid	Space
1. Which version enabled you to read more insightful comments?	16%	8%	75%
2. In which version are you more likely to leave your own comment or response?	16%	16%	67%
3. Which version would you prefer to use if you wanted to participate in a discussion about US politics?	8%	8%	83%
4. In which version do you expect to spend more time reading comments and browsing?	8%	16%	75%
5. Which version highlights the most insightful comments?	8%	33%	58%
6. In which version did you see more diversity among comments?	16%	33%	50%
7. Which version do you prefer overall?	8%	0%	92%

Table 4. Summary of responses to the exit survey, which asked participants to rank the interfaces after trying all three.

Participants indicated the degree of their agreement with a total of 782 comments (281 comments in the List interface, 249 in Grid, and 252 comments in Space) on a continuous scale from 0.0 (strongly disagree) to 1.0 (strongly agree). Average values are reported in Table 2. Bartlett's test on this data gives a p-value of 0.850 >> 0.05, suggesting that the homogeneity of variances assumption is valid. ANOVA yields a p-value of 0.00002073 << 0.05, and a follow up analysis with a two-tailed t-test shows statistical significance between all pairs of interfaces. P-values for each pair are: 0.03335 for Grid-Space, 0.000000149 for Space-List and 0.002115 for List-Grid, which supports H4.

Hypothesis 5 (H5): Opinion Space users will report significantly greater respect for the comments of others than they do when using the List or Grid interfaces.

Participants rated their degree of respect for a total of 782 comments by using a continuous scale from 0.0 (do not respect) to 1.0 (respect greatly). See Table 2 for the average values. ANOVA analysis yields a p-value of 0.001105 << 0.05, and a follow up analysis with a two-tailed t-test showed that users exhibited significantly greater respect for comments in both the Grid and Space interfaces as compared to the List interface (p-values are: 0.0007299 for List-Grid and 0.00003479 for List-Space). However, we did not find a statistically significant difference in respect values between the Grid and Space interfaces (p-value of 0.1191). We believe this is because both Grid and Space use the same visual method for highlighting the most insightful comments by adjusting the size and brightness of the points.

Comment Browsing Strategies

Participants were also asked to report how they selected comments to read in each interface. For the List interface, 6 participants replied that they read the comments in the order they were displayed, and the other half said that they randomly selected the comments.

For the Grid interface, 7 out of 12 people replied that they tried to diversify the comments they read by selecting a balanced combination of large and small point sizes. Four people said that they picked the points in random order and did not pay attention to the point size. Only one replied that she started with the biggest point size and continued in descending order of point sizes.

Survey responses for the Space interface are presented in Table 5. 11 out of 12 participants reported that their strategy for reading comments was to diversify by clicking on points positioned far from their own.

DISCUSSION AND CONCLUSION

Conventional list-based comment interfaces do not scale well: as the number of comments grows, users quickly become overwhelmed and read only a few comments, often the most recent or most extreme as voted by binary "thumbs up / down" ratings. We designed Opinion Space as a scalable way to visualize the "opinion landscape" and to operate as a self-organizing system that encourages participants to find and consider comments written by those who hold opinions different from their own.

We found that users were significantly more engaged with the Space and Grid interfaces as compared to List in terms of dwell time per comment, and participants perceived the Space interface to be significantly more engaging than Grid and List and indicated by subjective rankings of the three interfaces (H1). We also found that participants reported significantly greater agreement (H4) with the comments they read using the Space interface, and they had significantly more respect for comments they read using Grid and Space as compared to List (H5). Our hypothesis that users would find the Space interface significantly more conducive to finding useful comments (H2) was marginally supported. These results are consistent with the results reported by Ludford et al [15], where online participants in movie discussion groups were more engaged when the diversity of viewpoints and the uniqueness of each participant's opinion were conveyed.

Our hypothesis that participants using the Space interface would read significantly more diverse comments, based on Euclidean distance between responses to the profile statements (H3), was not supported by the data. However, as illustrated in Table 5, study participants describing their comment browsing strategies for the Space interface reported that they made use of the specific graphical layout and the position of their own opinion point to seek out comments written by those with a diversity of opinions.

Participant No	What strategy did you use to explore the comments in the space?	Participant No	What strategy did you use to explore the comments in the space?
1	Explored all the extreme opinions and ones very close to mine as well.	7	Random at first, to see what was there. Then I began looking at opinions in different areas of the space to see how those of different viewpoints thought about this particular issue.
2	I chose a circle on the left side, then chose a corresponding circle on the right side. Also, I started from the periphery and came in towards the middle.	8	I looked at the points that were nearest me and furthest from me just to see if the system was accurate.
3	I only picked a few near me...then I picked the ones farthest from me. And then I looked at the landmarks. Picked a few near big clusters	9	I picked a few comments near where mine were so I could just see what likeminded people though. Then I picked comments far away from mine to see what other people on the social/political/moral spectrum thought.
4	I first chose the ones by me. Then I chose the particularly brighter and darker points. I chose the brighter points because I assumed that they would be in conflict with mine. After that, I chose the darker points for the same reason; I assumed that they would be more aligned with my views.	10	I tried to pick a variety of points on the left and right and points that were bright and dim.
5	Random	11	I checked the politicians'/commentators' opinions first, then took a look at one of the points near mine, then at the farthest one I could find, and sort of hopped back and forth from there, looked at some around the large blue points, looked at some at random...
6	I tried to get a good cross-section of differing opinions and views so that I would be able to view and try to understand all sides of the argument.	12	I clicked on points that ranged from being very close to my position and very far.

Table 5. Comment browsing strategies for the Space interface as reported by study participants.

Comment diversity was also high with the List and Grid interfaces. The chronological ordering of comments in the List and Grid interfaces induced a random ordering of diversity (relative distances) between comments, so these interfaces were also effective on average for exposing participants to a diversity of comments. The outcome may have been different if the List interface had been sorted based on binary "thumbs up/down" ratings, which would highlight more extreme viewpoints. On the other hand, it is interesting and encouraging to note that the graphical display of Opinion Space did not significantly bias users toward only reading comments written by those with similar opinions.

FUTURE WORK

Although comment lists have many faults, they have one huge advantage: they are familiar to users. This user study suggests that Opinion Space can be effective, but our primary challenge is reducing the barrier to entry by making the interface easy to use and more intuitive. Opinion Space is a new model; its spatial arrangement of points may not yet be intuitive to users who expect to see the space labeled with axes such as "liberal" and "conservative." We view this as potentially a strong advantage – it conveys that the range of opinions do not fall along a single axis and that they are far more diverse. However, feedback we have received from users suggests that they want to better understand the arrangement. One idea we are exploring is to insert "landmarks", well-known people such as Jon Stewart or Oprah Winfrey into the space, and to automatically label regions of the space by clustering the points and performing textual analysis on the comments in each cluster to extract significant keywords that can be overlaid on the space.

We are also curious whether a scoring model can introduce incentives to increase user engagement. We posit that there are three types of users: 1) casual users who want to quickly find and read the most insightful comments, 2) "authors" who want to contribute eloquent comments that gain the respect of other participants, and 3) "gamers" who want recognition for their role in shaping the space by rating the comments of many others. We are developing new scoring metrics for these purposes, with close attention to avoiding malicious user behavior.

The user study reported here was limited to one hour per participant. To further investigate behavior over time, we would like to conduct a longitudinal user study. We are currently working with the U.S. Department of State to develop a version of Opinion Space that will solicit and highlight the most insightful ideas and viewpoints on U.S. Foreign Policy from a broad range of international participants.

We are now exploring how Opinion Space might be extended and applied to commercial websites such as Netflix, Amazon, Slashdot, and Digg. A scalable tool for managing massive online discussions requires a method for filtering user-generated data. In future versions of Opinion Space, we will extend our work on Eigentaste [12], a PCA-based collaborative filtering algorithm that runs in constant online time, and combine it with our model for identifying insightful comments [3] to make personalized comment recommendations.

ACKNOWLEDGMENTS

We are grateful to the additional members of the Opinion Space team: Tavi Nathanson, David Wong, Elizabeth Goodman, Gail de Kosnik, Alex Sydell, Ari Wallach, Christopher Goetz, Dhawal Mujumdar, Meghan Laslocky, Rupa Saheli Datt, Susan Miller, and Zach Blas. Many thanks to Warren Sack, Judith Donath, Henry Brady, and Bobby Nyotta, Maneesh Agrawala, Bjoern Hartmann, and John Canny for their insightful feedback. We also thank the organizations that support this research: the Berkeley Center for New Media, Jim Buckmaster of craigslist.com, and an NSF Graduate Research Fellowship.

REFERENCES

1. A.J. Berinsky. The two faces of public opinion. American Journal of Political Science, pp. 1209-1230, 1999.
2. J. Bishop. Increasing participation in online communities: A framework for human--computer interaction. Computers in Human Behavior, 23(4):1881-1893, 2007.
3. E. Bitton. A Spatial Model for Collaborative Filtering of Comments in an Online Discussion Forum. Proc. Of RecSys, ACM, 2009.
4. P.J. Carrington, J. Scott, and S. Wasserman. Models and Methods in Social Network Analysis. Cambridge University Press, 2005.
5. W.J. Conover, R.L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. American Statistician. American Statistician, 35, 124–129, 1981.
6. P.B. Brandtzaeg and J. Heim. Explaining Participation in Online Communities. Handbook of Research on Socio-Technical Design and Social Networking Systems, 2009.
7. P. Dahlgren. The Internet, public spheres, and political communication: Dispersion and deliberation. Political Communication, 22(2):147-162, 2005.
8. L. Freeman. Visualizing Social Networks. Journal of Social Structure, 1(1):4, 2000.
9. J.S. Fishkin and R.C. Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. Acta Politica, 40(3):284-298, 2005.
10. I.K. Fodor. A survey of dimension reduction techniques. LLNL Technical Report, UCRL-ID-148494 (2002).
11. M. Girvan and MEJ Newman. Community structure in social and biological networks. Proc. of the National Academy of Sciences, 99(12), 2002.
12. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval Journal, 4(2), pp. 133-151. July 2001.
13. J. Heer and D. Boyd. Vizster: Visualizing online social networks. In Proc. of the IEEE Symposium on Information Visualization, pp. 33-40, 2005.
14. I.T. Joliffe. Principal Component Analysis. 2nd ed. Springer-Verlag, 2002.
15. P.J. Ludford, D. Cosley, D. Frankowski, and L. Terveen. Think different: increasing online community participation using uniqueness and group dissimilarity. Proc. of ACM CHI, pp. 631-638, 2004.
16. S. McKillup. Statistics explained: an introductory guide for life scientists. Cambridge University Press, 2006.
17. B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2):1-135, 2008.
18. A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. IEEE Transactions on Visualization and Computer Graphics, 12(5):693-700, 2006.
19. L. Rourke and H. Kanuka. Barriers to online critical discourse. International Journal of Computer-Supported Collaborative Learning, 2(1):105-126, 2007.
20. W. Sack. Conversation Map: An Interface for Very Large-Scale Conversations. Journal of Management Information Systems, 17(3), 2001.
21. W. Sack. Discourse architecture and very large-scale conversation. Digital Formations: IT and New Architectures in the Global Realm, pp. 242-282, 2005.
22. C. Sunstein. Republic.com 2.0. Princeton University Press, 2007.
23. MJW Thomas. Learning within incoherent structures: the space of online discussion forums. Journal of Computer Assisted Learning, 18(3):351-366, 2002.
24. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.
25. F. Viegas and J. Donath. Social network visualization: Can we go beyond the graph. In Workshop on Social Networks, CSCW, volume 4, pages 6-10, 2004.
26. B.L. Welch. The generalization of "Student's" problem when several different population variances are involved. Biometrika 34 (1–2): 28–35, 1947.