The Impact of Social Information on Visual Judgments

Jessica Hullman, Eytan Adar
University of Michigan, School of Information
Ann Arbor, MI, USA
{jhullman, eadar}@umich.edu

Priti Shah
University of Michigan, Department of Psychology
Ann Arbor, MI, USA
priti@umich.edu

ABSTRACT

Social visualization systems have emerged to support collective intelligence-driven analysis of a growing influx of open data. As with many other online systems, social signals (e.g., forums, polls) are commonly integrated to drive use. Unfortunately, the same social features that can provide rapid, high-accuracy analysis are coupled with the pitfalls of any social system. Through an experiment involving over 300 subjects, we address how social information signals (social proof) affect quantitative judgments in the context of graphical perception. We identify how unbiased social signals lead to fewer errors over non-social settings and conversely, how biased signals lead to more errors. We further reflect on how systematic bias nullifies certain collective intelligence benefits, and we provide evidence of the formation of information cascades. We describe how these findings can be applied to collaborative visualization systems to produce more accurate individual interpretations in social contexts.

Author Keywords

Graphical perception, information visualization, social proof, social influence, Mechanical Turk.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Experimentation, Human Factors, Measurement.

INTRODUCTION

More individuals are generating and analyzing interactive data visualizations online than ever before, thanks to a growing number of social visualization sites like ManyEyes [43] and Swivel [40]. The need for such systems is in part driven by the ample availability of open data and the strong belief that collective analysis of this data might produce better understanding of this information. Social interaction, through visualization annotation, is suggested to be a primary motivator of use [20, 46, 44]. This leads researchers and system designers to consider how they might further increase users' engagement in these environments through design features that support and encourage social interaction. Such features include active and passive collaboration tools that range from threaded discussions, social embedding, tagging and other social annotation schemes. The most evident impact of these designs is that the visualization artifact is no longer considered independently of social content as prior members' responses and observations become attached to the visualization. This represents a significant shift from traditional visualization representation where any commentary or annotation was limited to a small group.

Online visualization communities may generate useful insights into data sets, under conditions that lead to the majority answer being more likely to be correct than any individual response [37, 39]. It is also possible that visualization users may misinterpret data when one or more prior users within the community have made errors in their interpretations of data. Of particular concern is the possibility of an erroneous information cascade in which initial members seed a discussion with inaccurate interpretations that get further distorted over time. A viewer new to a complex set of data with numerous options for creating visualizations may rely on the visualizations and interpretations that prior users have generated to constrain the search space. This is not unlike studies on patterns of bias arising from *social proof* in cultural markets like music downloading websites, where the popularity of artifacts becomes unpredictable and subject to a "rich get richer" dynamic [34, 17].

The principle of social proof—which suggests that actions are viewed as correct to the extent that one sees others doing them—falls under the larger category of social influence effects: those where a subject's feelings, thoughts, or behaviors are influenced based on observations of others' behavior in a similar situation. Studies of how social influence and conformity affect decision-making date back to the 1950's and earlier. In the context of visual perception, Asch's line experiment famously showed that the subject's responses for a simple length judgment task can be influenced by the answers first supplied by confederates [2]. Many new experiments have since identified different types of social influence and the contexts in which they are effective [9, 5, 49].

Interpreting a data visualization involves a complex set of cognitive and perceptual processes that have been identified by psychological research on graph comprehension (see [36]). Psychological models of graph interpretation focus on how the visual properties of a graph

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

(e.g., it's color, size, format, and so forth) influence how easily different information can be encoded from a data set (e.g., [11]). Although there is some evidence that prior beliefs and expectations have an influence on graph interpretation (e.g., [35]), the focus of research has been on an individual's own prior beliefs rather than the effect of beliefs of others. The work presented here seeks to extend graphical comprehension models to take into account social influences so that these models are appropriate for describing data interpretation in the context of social visualization systems. At the basis of any interpretation from visualization is an underlying graphical perception task. While we hope to eventually close the gap between basic graphical perception tasks and the higher-level interpretations that they lead to, our goal here is to provide initial evidence for the possibility for social influencers like social proof to occur in graphical settings.

We begin by modifying Cleveland and McGill's original experiments on the accuracy of visual judgment types (as executed by Heer and Bostock's Amazon Mechanical Turk implementation [19]). By including socially-derived signals (e.g., histograms of putative previous answers), and testing for other potential effects (e.g., anchoring), we are able to assess the impact of social influence by adjusting the bias in these signals towards or away from the "true" answers. Our work illustrates how social proof with an *unbiased* signal results in *more* accurate estimates_but that a *biased* signal results in *less* accuracy. Additionally, we extend the classical experiments to include a more difficult, and arguably more realistic task, that of judging linear associations between variables in a scatterplot. We also provide evidence that biased collective estimates for a graphical perception task can emerge in social environments through information cascades. We use the insights gained from our experiments to identify the implications for social visualization systems and discuss the impact of social proof and other forms of social influence on visualization research.

RELATED WORK

Social visualization environments have captured the attention of researchers in visualization, HCI, and CSCW who seek to understand the features of successful social data analysis systems. Heer, Wattenberg, Viegas, and others have completed a series of works aimed at describing the space [20, 43, 44, 46]. Motivated by work in various fields with socio-organizational bents, the authors of [18] present design considerations for asynchronous collaboration in visual analysis environments, highlighting issues of work parallelization, communication and social organization.

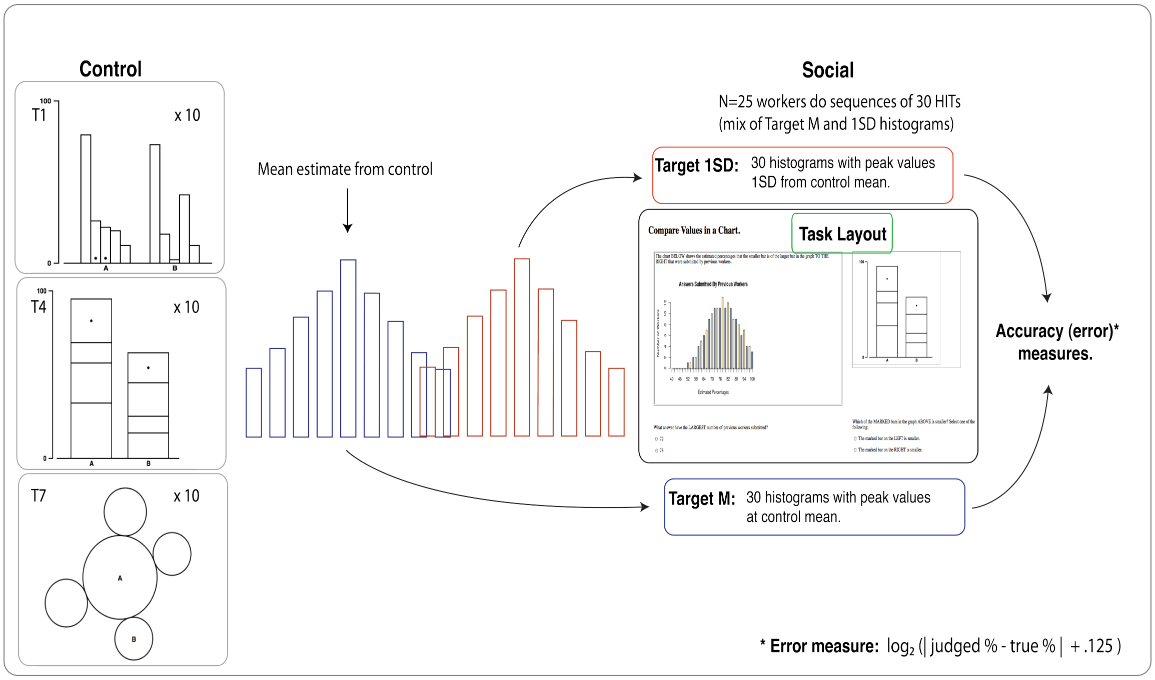
While highlighting many useful design considerations, the tone of these investigations is optimistic in that pitfalls that may stem from social processes, individual biases, and their combination are rarely considered. Instead, researchers discussing socially-related challenges tend to focus the most on the tendency toward imbalances in contributions among members [29]. There remains little work beyond Asch's classical experiments that attempts to combine perception and social influence tasks. We briefly summarize the two areas as they motivate our thinking and experimental design.

Graphical Perception

Graphical perception is a mainstay of visualization research [11, 41, 7, 45] due to its potential to improve the efficiency of automatic presentations of data [26]. Graphical perception can be affected by both design and data characteristics, warranting its continued investigation. A recent set of experiments executed by [19] demonstrated the use of Amazon's Mechanical Turk (MTurk) as a means for replicating prior studies and producing new knowledge in this area. Statistically comparable results are demonstrated despite the lack of control over screen resolution and other technical conditions. Qualification tests consisting of sample questions for a target task help control for workers' prior experience with graph interpretation and statistical literacy. For example, Heer and Bostock replicate one experiment from Cleveland and McGills's seminal study ([11]) to rank visual variables such as length, area, and color for encoding quantitative data. A proportion judgment task executed on bar, stack bar, and pie charts (among others), is used to rank the dominant visual variables on which the chart types are based. Heer and Bostock's results match the original authors' ranking of visual judgment types and add rankings for additional chart types. Though MTurk experiments do not address social influence, they do demonstrate a type of collective intelligence applied to perception tasks (see also [24]). As Heer and Bostock and others have suggested [33], Mechanical Turk offers a greater diversity of subjects, scalable experimentation, and rapid responses.

An important point with regard to our study of influence is that many graphical perception tasks are based in intuitive judgment, such as scanning a plot to form an impression of the underlying distribution. Intuitive judgments are typically faster, require less effort, and are less subject to over-thinking than analytic reasoning [1, 23]. They can benefit the perceiver of a graph, by displaying properties of the data that remain hidden when only the statistical parameters are computed [27]. However, they can also mislead an observer's interpretation of a stimulus [1, 23]. Hence, the fact that graphical perception results are replicable in some cases does not mean that subjects' answers are always correct. In some cases, systematic biases at the individual level may affect responses (see [36] for a compilation). Proportion-judging, for example, has been cited as one task where systematic biases can occur [38, 36]. In such cases, rather than errors distributing evenly in both directions from the actual answer, such biases may lead to potentially "bad" social signals.

Linear association estimation is another task subject to numerous biases and considered relatively difficult for humans [15]. Intuitive estimates have in many cases been found to be lower than the statistical coefficient *r* [28].


Control
T1
100
0
A
B
x 10
T4
100
0
A
B
x 10
T7
A
B
x 10
Mean estimate from control
Social
N=25 workers do sequences of 30 HITs
(mix of Target M and 1SD histograms)
Target 1SD: 30 histograms with peak values 1SD from control mean.
Task Layout
Target M: 30 histograms with peak values at control mean.
Accuracy (error)* measures.
Answers Submitted By Previous Workers
Number of Workers
0
2
4
6
8
10
12
40
46
52
58
64
70
76
82
88
94
100
Estimated Percentages
100
0
A
B

Compare Values in a Chart.

The chart BELOW shows the estimated percentages that the smaller bar is of the larger bar in the graph TO THE RIGHT that were submitted by previous workers.

What answer have the LARGEST number of previous workers submitted?

○ 72

○ 78

Which of the MARKED bars in the graph ABOVE is smaller? Select one of the following:

○ The marked bar on the LEFT is smaller.

○ The marked bar on the RIGHT is smaller.

* **Error measure:** $\log_2(|\text{judged \%} - \text{true \%}| + .125)$

Figure 1: Experimental Design

Further, estimates can be affected by manipulating visual characteristics, complicating the accurate judging of an association. While social data analysis from visualizations often involves further interpreting these underlying visual relationships, errors at the perceptual level undoubtedly play a role in determining the ultimate analyses.

Social Influence and Social Proof

Social psychologists use the term *social influence* to describe the tendency to respond in certain ways to the behaviors of others. For example, *conformity* refers to changing one's behavior to match responses of others. In analyzing conformity, social psychologists typically distinguish between *informational motivations,* arising from a desire to form accurate interpretations, and *normative motivations* deriving from a desire to obtain social approval [14]. This distinction is important in framing our work's focus on influence as applies to visualization perception to prior influence experiments that utilize perception tasks. Asch's well-known line judgment experiment [2] showed that individuals asked a simple visual judgment question (to identify which of three lines drawn on a blackboard matches a target line) responded differently depending on the answers first reported by other individuals in the room. The powerful support provided that elementary visual judgments can be subject to social influence inspired various replications and extensions (surveyed extensively in [5]). In the Asch paradigm, identifiability and social presence—normative social factors—characterize the setting [30]. However, subsequent experiments [5] suggest that a controversy with this experimental set-up is the difficulty of replication within different cultures or time periods. Our work avoids some of these pitfalls by concentrating on informational rather than normative influence.

Under conditions when others' opinions are expressed as quantitative estimates and "judge" and "advisor" are equally well-informed, averaging represents an optimal strategy for accurate estimation because it cancels out errors [37]. Yet in the case of social environments where multiple individuals view and interpret a visualization, can displaying information on prior responses lead to social signals that are biased, thus negating the effectiveness of averaging? Our work is motivated by exploration of the multiple ways in which social information might be represented in interactive visualizations, including comments, bar and pie graphs, and graphical annotations [48]. We make use of a histogram as a concise representation of social information to avoid the additional complexities of comments that might confound our experimental objectives.

While not targeted at information visualization, recent studies on the social dynamics of online cultural markets motivate some of the design of the present work. The concept of social or observation learning [3, 4] describes cases where an abundance of options leads to conditions where popularity is taken as a signal of quality. This theory has motivated experiments that simulate information cascades in social environments in which individuals motivated to make informed choice use social signals. Experiments on online music downloading markets [34] and collaborative tagging such as bookmarking [17] find that success in online cultural markets is difficult to predict but can be described through a stochastic urn model based on social influence signals restricted to information on others' behavior. Similarly, work in recommender systems provides evidence that recommendations can change users opinions, where users tend to rate toward the system's predictions [12]. With regard to online reviews [49], it is shown that exposure of previous opinions induces trend following and ultimately the expression of increasingly extreme views. The cost of expressing an opinion when other previous views are known tends to lead to a selection bias that softens the extreme views.

This form of influence, arising from information on the prior decisions of other community members, is similar to online collaborative visualization environments, where prior responses are commonly provided to the viewer (sometimes as text or numerical summaries and sometimes depicted with yet another visualization). As in the music downloading environments described by [34], the number of possible visualizations calls for a natural heuristic for dealing with the choice overload, and people may benefit through interaction, and notions of commonality when they coordinate their choices with others.

These results echo the concept of social proof, which states that a behavior is viewed as correct in a given situation to the extent that others are performing it, and that more people, more ambiguous situations, and an increased sense of similarity to others increases the power of the influence [8].

STUDY OBJECTIVES

For this work we conducted a number of large-scale experiments using Amazon's Mechanical Turk (as well as additional validation experiments). We specifically tested the following main hypotheses:

- H1: Adding a social information signal on prior workers' behavior (responses submitted for a graphical perception task) will directly influence subject's accuracy on the task at hand.
  - H1a: If the social signal is an accurate representation of the true answer (i.e., unbiased), errors will decrease for those witnessing that signal.
  - H1b: If the signal is inaccurate (i.e., biased), errors will increase.
- H2: Biased responses can emerge through information cascades and initial conditions
  - H2a: Accuracy will be significantly affected, in the same direction as the social signal.
  - H2b: For an increase in the number of people included in the social signal (*n)* there will be a concurrent increase in the weight of the social signal on the judgment of person *n+1*.

Below we describe the main experiments used to test these hypotheses. The first extends Cleveland and McGill's seminal study of visual judgment types by examining how social proof affects visual judgments. The second, which adds another layer of task difficulty, is informed by work in the estimation of linear associations [10, 32, 28]. We then describe an experiment for simulating an information cascade in order to determine the likelihood that biased collective estimates for a graphical perception task might to naturally emerge in online social environments.

EXPERIMENT 1: PROPORTION JUDGMENT

Method

Control. Our first experiment was designed to ascertain the impact of social information on classical graph perception tasks. As a control, we began by adapting Heer and Bostock's 1a experiments [19] using Mechanical Turk. We limited our replication to three chart types distributed across their reported ranking of visual judgment types: a bar graph representing the high accuracy encoding type position along a common scale (T1), a stack bar graph representing primarily length encoding (T4), and a bubble chart representing relatively low accuracy circular area encodings (T7). We depict these within a representation of our experimental design (Figure 1).

Workers who accept the task examine a graph such as a bar or pie chart where two bars or sections of the graph are marked and then answer two questions: 1) "Which of the two MARKED bars is larger?" and 2) "Make a quick visual judgment on what proportion the smaller is of the larger." Like Heer and Bostock, we used the first question to verify responses and required workers to first pass a qualification test of several examples with multiple-choice questions. Each of the 30 unique chart versions was launched as an individual HIT to be completed by 50 workers for a reward of $0.05. This was raised to $0.08 to increase completion speed [27].

Social Conditions. Using the control data we implemented two social conditions through a social signal. Specifically, a *social histogram* showed a distribution based on 50 previous answers for the same judgment task. For each chart of the 30 charts, two social histograms were generated: Target M was set to the mean answer found in the control, and Target 1SD was set to one standard deviation from the mean in the direction of the greatest density of the control distribution (ranging from 3.15 to 14.8). We chose this particular target in order to test the case where the more incorrect social information might still be relatively believable. In all situations, Target M means were closer to the true proportion values than Target 1SD. In generating the histograms, we required that the value with the highest count in the histogram fall within three of the control mean used to generate the sample. The task required the worker to first identify 1) "What answer the LARGEST number of previous workers submitted?" followed by the two questions in part 1A. The task layout can also be seen in Figure 1.

Because participants saw a mix of biased (Target 1SD) and unbiased (Target M) histograms, one possibility was that whatever version they saw in their first HIT would influence subsequent HITs. Stated another way, if a biased histogram was shown first, was the subject less likely to rely on the social signal for subsequent HITs? To insure that a worker's judgments were not affected by ordering, we ran several pilot experiments consisting of three chart

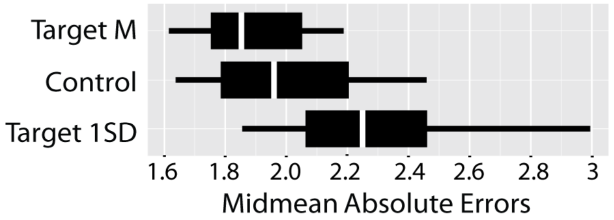


Figure 2: Mean log absolute error measures for control, Target M, and Target 1SD conditions of proportion task.

tasks per HIT with half the 50 subjects first seeing the Target 1SD histogram and the other half the Target M (with the second and third charts randomly assigned). We saw no statistical difference (p > .05) between perception errors in the subsequent charts, indicating that whether the first observed social histogram was accurate or inaccurate does not appear to impact how subsequent histograms were perceived.

We launched the 60 chart/histogram combinations in sets where each worker could do 30 unique chart tasks, with equal numbers of Target M and 1SD histograms and HIT order counterbalanced by the Mechanical Turk. All workers taking part in the social condition completed a qualification task, as before, with the addition of several example and practice histogram reading tasks, presented separately from the proportion judgment portion of the qualification.

Results

Control. A small majority (56%) of the workers accepted all 30 available HITs. We included in analysis all workers who had completed HITs regardless of the number (sensitivity analysis described below). We used the verification question to exclude workers with incorrect responses. Because we were concerned with the bias that outliers might lead to in the worst case scenario of charts with a large standard deviation in responses, we removed outliers by defining a range around the actual proportion for each chart (+/- 40) based on approximately 3x the largest standard deviation for a chart and omitting values outside of this range (a total of 6.0% of responses removed as outliers). To validate our control experiment, we use the midmeans of log absolute errors (i.e., the mean of the middle two quartiles or MLAE) using $\log_2(\ |judged\_percent - true\_percent| + .125)$ for each chart and bootstrapping (following [11]). We find that the ranking of judgment types is preserved and that the rough shape and ranking of visual judgment types by accuracy are preserved (relative to [19] and [11]).

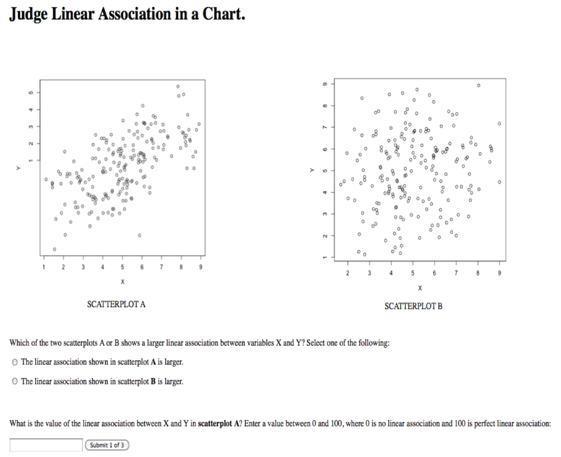
Social Conditions. An average of 22% of workers completed all 30 tasks in a sequence. We again considered all HITs for analysis regardless of the number done by individuals, removing outliers and those who didn't understand the task according to the procedure for the control. We also excluded HITs where the histogram verification question was answered incorrectly. A total of 7.2% of HITs were removed. As above, we calculated the MLAEs for each of the 60 unique chart/histogram pairs. Ignoring the particular chart type, we grouped all calculated MLAEs by experiment (i.e., control, Target M and Target 1SD). Figure 2 shows boxplots of the means from the control data and each of the social conditions. After doing an ANOVA (p < .001) we used a Tukey HSD test to compare the MLAEs across all three conditions. We found a significant difference for Target M and 1SD (p < .001). The lowest errors appear in the Target M condition (mean: 1.886, stdev: 1.210), followed by the control (mean: 1.989, stdev: 1.197), then Target 1SD (mean: 2.267, stdev: 1.173). A significant difference also exists between Target 1SD and the control (p < .001), although not between the Target M condition and the control (p = .1792). These results are consistent with our hypothesis 1 (the further the social signal, the less accurate the estimate).

Sensitivity Analysis. To control for the mix of between- and within-subjects data we conducted two sensitivity analyses: collapsing errors by individual by condition (reanalyzing the difference between the Target M and 1SD conditions while controlling for within-subjects variance by computing average mean error scores for the M and 1SD conditions for each individual), and controlling for effects of particular individuals on results by re-running the ANOVA and Tukey test using a systematic leave-one-out design. In both cases, final t-tests yielded significant results (p-values < .05) for all but the Target M and control comparison.

EXPERIMENT 2: LINEAR ASSOCIATION ESTIMATION

Method

Control. In a second two-part experiment similar to 1, we pseudo-randomly generated 30 correlation values to use in generating scatterplots (10 unique values under .5, 10 values between .5 and .8, and 10 values over .8). We chose these bins after the suggestion ([47]) that many statisticians see |*r*| values below 0.5 as "small", and values of |*r*| as "large" only when they are 0.8 or greater. For each correlation we generated 100 pseudo-random x values from a Gaussian distribution as well as pseudo-random y values for each. We transformed these (x, y) pairs by adapting [32]'s method to generate only positive linear associations, resulting in a final 30 scatterplots, one depicting each of our



Judge Linear Association in a Chart.

Which of the two scatterplots A or B shows a larger linear association between variables X and Y? Select one of the following:

○ The linear association shown in scatterplot **A** is larger.

○ The linear association shown in scatterplot **B** is larger.

What is the value of the linear association between X and Y in **scatterplot A**? Enter a value between 0 and 100, where 0 is no linear association and 100 is perfect linear association:

Figure 3: Layout for linear association task.

original correlations. In the individual tasks, the worker is shown a labeled scatterplot of the two variables X and Y and asked to estimate the value of the linear association between them on a scale of 0 and 100, the scale chosen after [10]. As in experiment 1, the 30 unique plots were launched as 30 individual HITs with N=50 assignments.

Each HIT (see Figure 3) consisted of two radio button verification questions: 1) "Which of two scatterplots [A and B shown side by side in the HIT] shows a larger linear association between variables X and Y?", and 2) "Would you describe the linear association in scatterplot A as high, medium, or low? Assume high is over 80, medium is from 50 to 80, and low is less than 50." The workers were then asked to enter a value between 0 and 100 describing the linear association in scatterplot A, where 100 represented perfect linear association and 0 represented no association. Each worker first took a qualification test showing examples of 100 and 0 association plots plus two additional examples and three practice tasks. Due to the relative difficulty of accurately estimating correlations, we allowed subjects to pass the qualification given 11 out of 12 correct answers. The B scatterplots used for comparison to the A plot in the first question of each HIT had an average difference of 61 from the true value of the A scatterplot.

Social Conditions. We modified the control as in experiment 1, adding two social influence conditions for each of the 30 scatterplots, where the target guess used to generate the distribution for the histograms of previous answers is the actual mean guess in one condition (M) and one standard deviation in the other (1SD) which ranged in practice from 3.07 to 22.2. In order to keep the presentation of the scatterplots A and B the same as in the control, we presented the histogram and histogram question beneath the two plots and question 1. Question 3, which asked the worker to estimate the association, was below the histogram and question 2. The 60 chart/histogram pairs were launched as individual hits in sets of up to 30 unique charts, with equal numbers of M and 1SD histograms and order counterbalanced between workers.

Results

Control. Under this condition, 75% of the workers accepted all 30 HITs in a sequence, yet we considered all completed HITs for analysis (sensitivity analysis described below). We used the verification questions to validate that subjects understood the task, excluding HITs with one or more wrong answers. We also excluded from analysis outliers that were more than 50 off from the actual correlation in the scatterplot (a total of 7.9% were removed), defining this boundary using approximately 3x the largest chart standard deviation. For each of the scatterplots, we calculated the mean estimated linear association and standard deviation.

We compared the pattern of results from our experiments to those of prior work in linear association estimation [10, 32]. Figure 4 shows the actual pattern observed (with a hand drawn line, in red, illustrating the expected patterns given previous literature that demonstrated maximum errors at "medium" correlations). We first noted that accuracy, as defined by the log absolute error of estimates, dipped at linear association levels of 50 and 80, then jumped back to the more expected trend. We attribute these jumps to the anchoring effect of the question text on the page. Recall that 50 and 80 were used to describe the transitions from low to medium to high correlations. Because this may have led to a slight bias in utilizing these answers, when the correlation was in fact 50 or 80, error was reduced. In future experiments we hope to eliminate these types of signals. However, this is not critical for the current experiment as we are not seeking to compare errors by task difficulty and data is aggregated for all correlation types. Additionally, outside of this difference, our pattern of log absolute errors accuracy measures matched prior results, in that the accuracy of estimates of association declined as the actual linear association moved closer to 50.

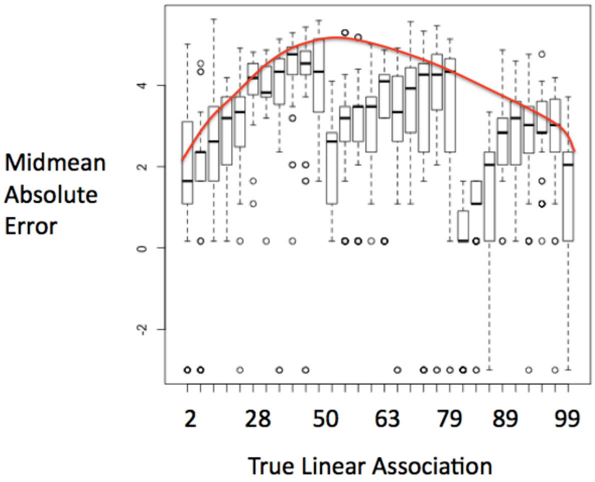


Figure 4: Log absolute errors across 30 linear association tasks (pre-bootstrapping), with expected pattern in red.

Social Conditions. An average of 58% of the workers accepted all 30 HITs in a sequence though we again considered all HITs in analysis, removing those that qualified as outliers using the procedure described for the Control. We also excluded HITs where the histogram verification question was answered incorrectly (a total 10.8% was removed between the two conditions). We calculated the midmean log absolute error accuracy measure as in Experiment 1 (including bootstrapping) for each of the 60 unique chart/histogram pairs, and ran an ANOVA (p = 0.4506). In this case, the log absolute error means for the three conditions were quite close (Target 1SD mean: 3.010, stdev: 1.122; control mean: 3.040, stdev: 1.128; Target M mean: 3.319, stdev: .8436). The lack of significance of the ANOVA and the relatively high errors from the Target M condition led us to consider the assumptions behind the hypothesis that as a social signal becomes more biased, errors will increase.

If we assume that the Target M histogram value is always between the actual answer for the task and the Target 1SD value (as it was in Experiment 1), then this pattern goes against our hypothesis that as a social signal

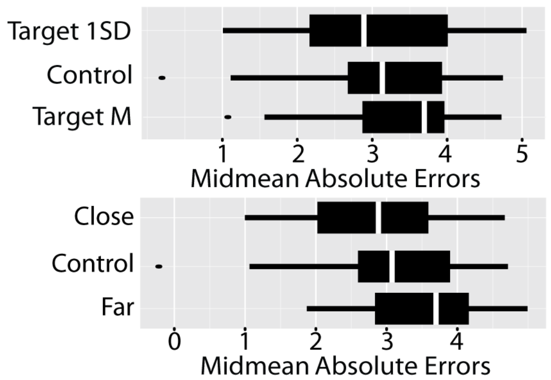


Figure 5: Mean log absolute error measures for control, Target M, and Target 1SD conditions of linear association task (above), and regrouped Farther and Closer conditions.

becomes more biased, errors will increase. In looking closer at the data, however, we realized that the results remained unclear with regard to H1, because in 16 of the 30 cases, the Target 1SD histogram value was in fact closer to the truth. Considering that prior research has shown humans to be relatively bad at linear association estimation, this outcome is plausible. Yet because the ordering is random, this confuses our relationships.

To overcome this complication, we regrouped the data based on which of the two histograms displayed a mean answer that was closer to the actual association (which we re-termed the Closer condition) versus the histogram that was farther from the actual association (Farther condition). With this measure, our previous observation—that the farther the social signal is from the actual, the less accurate the estimate—holds. An ANOVA yielded significance ($p < .01$). Tukey's HSD test showed no significant difference between the Control and Closer condition ($p = 0.5391$), nor between the Control and Farther condition ($p = 0.1113$). However, we saw a significant difference between the Closer and Farther condition ($p < .01$). Figure 5 depicts the ordering of mean absolute errors (Close mean: 2.767, stdev: .9533; control = 3.040, stdev: 1.128; Far mean: 3.561, stdev: .7740). We discuss these results further in the discussion section.

ANCHORING EFFECTS

Before going further to investigate the possibility that "bad" social signals might naturally emerge in a social visualization environment, we validate that the results of experiments 1 and 2 did in fact represent social influence rather than non-social influences. As examples of the latter, anchoring and adjustment [42] are psychological heuristics that subconsciously influence the way people intuitively assess probabilities. A subject starts with an implicitly suggested reference point (the "anchor"), and makes adjustments to that number based on additional information. Anchoring, which is related to priming, is the general activation of a particular idea or ideas based on associations between a stimulus and that idea(s).

To test for anchoring we ran a validation experiment that displayed the social histograms as in Experiment 1 but labeled them as something unrelated to the chart (which was itself labeled). For example, the histogram might be labeled "Temperature Recorded at Location 11" and a bar chart as "Employee Salary in Company R" (a set of unrelated labels that nonetheless made sense in the context of histogram and the different chart types were manually selected). This condition also made a clear delineation using different colored backgrounds behind each sub-task in the single HIT. We ran the conditions as 30 HIT sequences for $.10 a HIT with N=25 unique workers who had not yet done any of our prior tasks.

Prior to this we used a paired t.test to confirm that adding a label to the proportion-judging chart itself would not affect responses, by rerunning our control experiment with (N=25) but with a label (e.g. "Salaries of Employees in Company R" for a bar chart) above the chart ($p = 0.1214$).

Under the social condition with delineation, 72% of workers accepted all 30 HITs in the sequence (2.4% of responses were omitted in processing the results). After performing an ANOVA ($p < .05$), we used Tukey's HSD test to find that the delineated Target 1SD MLAEs were not significantly different than the control ($p = .1594$). As the same non-delineated Target 1SD histograms *were* different from control in the social task ($p < .001$), we can infer that anchoring is likely not contributing to increased errors as the histogram shifts from the "true" value.

EXPERIMENT 3: INFORMATION CASCADE AND INITIAL CONDITIONS

Method

Although the previous experiments clearly demonstrate that *if* an individual is presented with social histograms their judgment will change, it is not entirely obvious that *different histograms* would emerge from the *same social process* (e.g., an information cascade or initial conditions). Stated another way, we seek to establish whether judgments of the $n+1^{th}$ person are influenced by the number of previous judgments (n) and the distribution of those judgments. If an individual over-utilizes the judgments of others, one might expect that a) the initial condition would impact all subsequent judgments (e.g., if the first person is really off, everyone after them will be really wrong) and that b) the more individuals providing the estimate, the more "trusted" that social signal would be (e.g., person 31 relies more on the estimate than person 5).

If such a cascade pattern holds, an initial bad estimate may grow or become more entrenched as more and more people contribute their estimates. To simulate an iterated process we presented 1500 HITs that displayed histograms with varying n's and displaying a histogram with different means. An indication informing the participant that they were person $n+1$ in a series of individuals was made in 3 places on the interface (the histogram caption, question text, page heading, all in bold). Histogram means were centered at one standard deviation to the left (based on the control data), one standard deviation to the right, and at the mean of the control experiment (yielding 3 charts). The *n* variable was varied from 1 to 37 in steps of 4. Note that the control experiment serves as a test at *n*=0 (i.e., the participant is the first to make a judgment). In total, we produced 30 histograms per chart. This was done for all 10 circular area plots (chart type T7) yielding 300 total variants, which we launched as 10 HIT sequences at $.08 per HIT with 5 workers per variation (total of 1500 HITs).

We constructed a number of linear models to test for the relationship of actual answer (the Turker's judgment) against histogram mean, *n*, and the true proportion. Specifically, the model:

$$Answer_{n+1} = b_0 + b_1 * true\_proportion + b_2 * histogram\_mean + b_3 * histogram\_mean * n + e$$

attempted to capture the increasing effects of *n* on the answer as well as the participant's personal evaluation of the true proportion. Sensitivity analyses with robust standard errors were performed using generalized estimating equations in *R* to account for the repeated measures per person. These findings were robust to control for the correlation induced by collecting multiple measurements per study subject.

Results

Modeling the main effects of the actual value and the suggested histogram mean on the answer produced an adjusted r^2 of .8122. Both the actual value and histogram mean were positively associated with the person's answer with a slightly large effect for the true proportion (effect=0.65, standard error = 0.27, p < .001) than the histogram mean (effect=0.439, standard error = .033, p < .001). Interestingly, we could find no significant effect of *n* in this model or any other we used for sensitivity analysis.

The results of this experiment suggest that hypothesis H2b does not hold. In other words, the opinion of 5 people counts the same as that of 30, and the first judgment, erroneous or not, sets the stage for all subsequent answers. For the sake of completeness we discuss below several potential limitations of our experimental conditions that might partially affect the results.

DISCUSSION

Findings and Implications

The main finding in this work is the evidence we provide that responses to online graphical perception tasks can be subject to influence from socially-derived information signals such as social proof via prior responses, and that such biased signals are possible given a situation where any *n+1* person can see the responses of the *n* individuals who saw a graph before him or her.

Social Errors. In the proportion judgment experiments, we observed a clear difference between the individuals exposed to Target 1SD histograms over Target M. Furthermore, because biased histograms like our Target 1SD might emerge from a cascade process there is a need for any system utilizing such social signals to be highly aware of this possibility, and to potentially mask social signals in situations where this type of bias might happen.

Systematic Bias. To mitigate information cascades, a designer might have the intuition to mask the social signal (e.g., hide the histogram) until a sufficient number of samples is obtained. However, as demonstrated by the lack of significant difference between Target M and the control, there does not appear to be any benefit (or conversely, harm) in displaying this information. This result is only surprising if one assumes that graphical perception lacks systematic bias. However, because individual judgments are wrong, and generally wrong in the same "direction," the overall collective opinion does not appear to be any better than the individual one. Worse, in situations such as the linear estimation task, when systematic bias and estimation errors were so high that Target 1SD histograms were equally likely to be closer to the true correlation as Target M (16 of the 30), there is a clear indication that the social signal, even the individually-derived one, has a negative impact on perception. This observation indicates that caution is necessary—or at least awareness—when designing systems in domains with systematic bias.

Social Influence and Future Work

Social influence, construed more broadly, is known to be a result of multiple features. In this study we have targeted a specific type of social influence, one centered around social proof, which we believe can serve as a jump-off point for future work.

One opportunity for future work is to address the reasons why, as in our Experiment 3, the number of previous responses (*n*) did not impact the model. Given that the results of our social conditions did appear to be based on the socially-derived signal, we would expect that more responses would result in a stronger social information signal, exerting a greater effect. Effects of this sort have been validated in other contexts by [34, 12]. Yet this was not suggested by our data. We hypothesize that it may be that the MTurk environment did not support the type of systematic processing that may have been required for a worker to sufficiently understand the relationship between the total count in the histogram and the potential value of the social signal. Utilizing a true iterated experiment (e.g., through [25]) may yield a different result. In addition, while we chose a one standard deviation difference in order to investigate cases where the social signal remains believable, future work might offer more insight into how biased a social signal can be while still exerting similar effects.

We also note that our study is not designed to induce normative social influence stemming from a desire for social approval. The histograms are informational in nature. Because the decisions of workers in our experiments were not witnessed in the presence of others, workers may have felt more confident in deviating from the distribution. While some recent research suggests that anonymity need not always degrade social influences effects in computer-mediated environments, it is suggested that a sense of social identity must be in place for influence to occur [31].

Furthermore, social influence is also known to be stronger when signals come from others whom the subject deems similar to her/himself [8]. The extent to which a person identifies with message source (majority or minority) is a significant factor in determining information processing strategies plus outcome of influence attempt [13]. Such theories indicate that stronger influence might be achieved had the presence of the other workers and their similarity to the subject been emphasized. Because users of social visualizations sites are not commonly designed around anonymity, additional normative effects may lead to more significant effects on judgments within these frameworks.

Another future inquiry might further investigate the effects of task difficulty on influence, as the relative difficulty has been cited to have an effect on the degree to which people accept advice [16]. Because the distance between the Target M and 1SD histograms in our experiment was defined relative to the variance of the control data (e.g. standard deviation), our results did not allow us to cleanly analyze whether task difficulty affected the level of influence. Yet such knowledge would offer designers of social visualization systems further insight into the particular types of situations where the risks of social influence are most heightened.

The contributions we make to social visualization system design are based on experiments that focus on a narrow type of task and environment. However, our results indicate that graphical perception, a key first step in the interpretation of visual information, can be influenced by social signals that may be present in collaborative visualization systems. Clearly, actual systems like ManyEyes are complex environments. Factors such as expertise and interest in the content, prior experience with graph interpretation and statistical literacy (see [36] for others), undoubtedly play a role in such systems. Such environments present individuals with graph comprehension tasks of varying difficulty, and may also present situations that fall along a continuum of objectivity with regard to the pattern being visualized. For example, while the tasks we investigate here have objective answers in the true proportions and linear associations that are visualized, there are many tasks where an objectively true answer might not be possible, such as graph aimed at visualizing an evaluation formed by subjective sentiments on a topic. These may still serve as important points of discussion and collective analysis, and thus potential distinctions in social influence patterns as determined by objectivity may offer further insight for designers.

CONCLUSION

In this paper, we have presented evidence suggesting that responses to graphical perception tasks online may be subject to social influence. We demonstrate through a large-scale study that social proof has an impact on visual judgment, and with it, perceptual accuracy. By modifying classic graphical perception tasks on proportion judgment and linear association estimation, we found a clear indication that an erroneously biased social signal will result in more errors and conversely that a less-biased one will lead to fewer errors. However, we also identify that systematic bias makes many socially derived signals (e.g., a histogram of individually-collected results) erroneous on the whole, and these signals do not commonly provide a definite benefit over individual assessments. This calls into question some of the benefits of "collective intelligence" and highlights a number of design risks. We also identify that initial seeds in social signals (e.g., the first person to contribute to the histogram) allow information cascades to rapidly take hold and impact all future answers. Collective visualization systems hold a great deal of promise for the great influx of data experienced today. However, previous work on visualization systems frequently ignores social effects, treating visualization interpretation as an individual process. As our study highlights, there is a need to form new theories and models that explain the impact of social processes on community-driven visualization environments and lead to new systems.

ACKNOWLEDGEMENTS

We thank Jeff Heer and Mike Bostock for making their materials and experimental details available. We are grateful to Sara Adar, Tom Finholt, Paul Resnick, Eytan Bakshy and Erin Krupka for their comments and advice.

REFERENCES

1. Alter, A.L., Oppenheimer, D.M., Epley, N., and Eyre, R.N. Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *J. of Exp. Psyc.-General 136*, 4 (2007): 569–576.
2. Asch, S.E. Effects of group pressure upon the modification and distortion of judgment. In *Groups, leadership and men*. Edited by H. Guetzkow. Pittsburgh, PA., 1951.
3. Banerjee, A.V. A Simple Model of Herd Behavior. *The Quarterly J. of Economics 107*, 3 (1992): 797-817.
4. Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashions, customs, and cultural change as information cascades. *J. of Pol. Econ.*, 100(5): 992–1026.
5. Bond, R., and Smith, P. Culture and Conformity: A Metanalysis of Studies Using Asch's (1952b, 1956) Line Judgment Task. *Psych. Bulletin* 119, 1, (1996): 111-137.
6. Bresciani, S., Blackwell, A. F. and Eppler, M. "A Collaborative Dimensions Framework: Understanding the mediating role of conceptual visualizations in collaborative knowledge work," in *HICSS*, (2008),: 364.
7. Card, S.K., Mackinlay, J., and Shneiderman, B. *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann, 1999.
8. Cialdini, R.B. *Influence: Science and Practice*. Allyn & Bacon, 2000.
9. Cialdini, R. B., and Goldstein, N. J. "Social Influence: Compliance and Conformity," *Ann. Rev. of Psych.* 55, no. 1 (2, 2004): 591-621.
10. Cleveland, W. S., Diaconis, P., and McGill, R. "Variables on scatterplots look more highly correlated when the scales are increased," *Science* 216, no. 4550 (1982): 1138–1141.
11. Cleveland, W.S. and McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Statistical Assoc.*, 79, (1984): 531-554.

12. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., and Riedl, J. Is seeing believing?: how recommender system interfaces affect users' opinions. *CHI'03*, (2003): 585-592.
13. David B, Turner JC. Majority and minority influence: a single process self-categorization analysis. In *Group Consensus and Minority Influence: Implications for Innovation*, ed. CKW De Dreu, NK De Vries,. Malden, MA: Blackwell (2001): 91–121.
14. Deutsch, M., & Gerard, H. B. A study of normative and informational social influences upon individual judgment. *The J. of Abnormal and Soc. Psych*. 51 (3, 1955): 629-636.
15. Doherty, M., Anderson, R., Kelley, A., Albert, J. Probabilistically valid inference of covariance from a single x,y instance when univariate characteristics are known. *Cog. Sci.*, 33 (2009): 183-205.
16. Gino. F. and D. Moore. "The Effects of Task Difficulty on Use of Advice." *J. of Behav.Decision Making*, 20 (2007): 21-35.
17. Golder, S. A. "Usage patterns of collaborative tagging systems," *J. of Info. Sci.* 32, no. 2 (4, 2006): 198-208.
18. Heer, J., and Agrawala, M. Design considerations for collaborative visual analytics. *Information Visualization, 7*, 1 (2008), 49-62.
19. Heer, J. and Bostock, M. "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *SIGCHI'10*, (2010), 203–212.
20. Heer, J., Viégas, F.B., and Wattenberg, M. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Communications of the ACM* 52, 1 (2009), 87–97.
21. Henrich, J., Heine, S.J., and Norenzayan, A. The Weirdest People in the World? SSRN eLibrary, (2010).
22. Kittur, A., Suh, B., and Chi, E.H. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. *CSCW '08,* (2008): 477–480.
23. Klein, G. *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*. MIT Press, 2009, 76-100.
24. Kosara, R. and Ziemkiewicz, C. Do Mechanical Turks Dream of Square Pie Charts? *BELIV '10* (2010): 373–382.
25. Little, G. TurKit: Tools for iterative tasks on mechanical turk. , *IEEE VL/HCC'09* (2009): 252-253.
26. Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Trans. Graph. 5*, 2 (1986): 110-141.
27. Mason, W., and D. J. Watts, "Financial incentives and the 'performance of crowds," Workshop on Human Computation *SIGKDD '09* (2009): 77-85.
28. Meyer, J., Taieb, M., and Flascher, I. Correlation estimates as perceptual judgments. *J. of Exp. Psych., Applied*, 3 (2005): 3-20.
29. Noel, S. and Lemire, D. On the Challenges of Collaborative Data Processing. 0906.0910 (2009).
30. Postmes, T, Spears, R, Sakhel, K, De Groot, D. "Social influence in computer-mediated communication: The effects of anonymity on group behavior," *Personality and Soc. Psych. Bulletin* 27, (10, 2001): 1243.
31. Reicher, S.D, Spears R., Postmes T. A social identity model of deindividuation phenomena. *Eur. Rev. Soc. Psychol.* 6 (1995): 161– 98.
32. Rensink, R., and Baldridge, B. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29, (2010): 1203-1210.
33. Ross, J., Irani, I., Silberman, M. Six, Zaldivar, A., and Tomlinson, B. (2010). "Who are the Crowdworkers?: Shifting Demographics in Amazon Mechanical Turk".*CHI EA '10* , (2010): 2863-2872.
34. Salganik, M. J., Dodds, P. S., and Watts, D. J. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science* 311, no. 5762 (2, 2006): 854-856.
35. Shah, P., and Freedman, E. (in press). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*.
36. Shah, P., Freedman, E., & Vekiri, I. The comprehension of quantitative information in graphical displays. In P. Shah and A. Miyake, (Eds.). *The Cambridge handbook of visuospatial thinking*. New York: Cambridge University Press, (2005): 426-476.
37. Soll, J. B., & Larrick, R. P. (2004) Strategies for revising judgment: How, and how well, do people use others' opinions? Unpublished manuscript.
38. Spence, I. Visual Psychophysics of Simple Graphical Elements. *J. of Exp. Psych.: Human Perception and Performance 16*, 4 (1990): 683-692.
39. Sunstein, C. R. *Infotopia: How Many Minds Produce Knowledge*, annotated edition. Oxford University Press, USA, 2006.
40. Swivel, Inc. (2007). Swivel. Retrieved September 18, 2010, from http://www.swivel.com.
41. Tufte, E. R. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Conn., 1983.
42. Tversky, A. and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science 185*, (4157, 1974): 1124-1131.
43. Viegas, F. B., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. 2007. ManyEyes: a Site for Visualization at Internet Scale. *IEEE TVCG* 13, (6, 2007), 1121-1128.
44. Viegas, F.B., Wattenberg, M., McKeon, M., Van Ham, F., and Kriss, J. Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. *HICSS '08,* (2008): 159.
45. Ware, C. *Information Visualization, Second Edition: Perception for Design*. Morgan Kaufmann, 2004.
46. Wattenberg, M. and Kriss, J. "Designing for social data analysis," *IEEE TVCG* (2006): 549–557.
47. M. B. Wilk. Bell Laboratories Technical Memorandum. Bell 1aboratories, Murray Hill, N.J., 1966.
48. Willett, W., Heer, J., and Agrawala, M. Scented widgets: Improving navigation cues with embedded visualizations. *SIGCHI '07*, (2007): 51-58.
49. Wu, F. and Huberman, B. How public opinion forms. *Internet and Network Economics*, (2008): 334-341.