Evaluating Approaches to Crowdsourced Visual Analytics

JESSICA HULLMAN, University of Washington
ERIN KRUPKA, University of Michigan
EYTAN ADAR, University of Michigan

1. INTRODUCTION

Information visualizations are used by various organizations to present data for analysis to a crowd, for example, to identify interesting patterns [Heer 2007], [Willett et al. 2012]. Visualization is an efficient means of presenting data to citizen analysts because well designed visualizations can leverage human information processing abilities [Larkin and Simon 1987].

Visualization crowdsourcing experiments often follow one of a limited number of workflows for distributing evidence. In a *one-to-many* workflow each crowd member is presented with the same single visualization and asked for a summary judgment (e.g., [Willett et al. 2012]). A challenge with this approach is that it does not encourage viewers to consider factors that *ought* to effect their assessment (e.g. variance resulting from a small sample size; [Tversky and Kahneman 1974]). Techniques for conveying variation, on the other hand, are often hard for individuals to understand [Belia et al. 2005]. Aggregating judgments made by the crowd about any particular feature (e.g., the presence of an outlier or cluster) results in a proportion describing how many crowd members recognized the feature. However, when crowd members are not sensitive to variance in the data, the aggregated judgment may be unduly biased to spurious patterns in the input dataset. In a *many-to-many* workflow, many crowd members have access to many visualizations depicting subsets of the same dataset (e.g., [Wattenberg and Kriss 2006], [Heer 2007]). Different graphical views of a single dataset may result in varying perceptions of the data (e.g., a non-optimal scaling applied in one view is absent from another) and lead to a greater sensitivity among crowd members to multivariate relationships. However, final assessments may be impacted by contextual biases such as the order in which the different visualizations are viewed. Crowd members who may still overlook factors like variance in examining any single view, resulting in bias in aggregated judgments toward the particular data instantiation.

We consider an alternative workflow that generates many datasets through bootstrapping and distributes visualizations of these resamples. Taken as a set, the resample datasets reflect variance in the input dataset such as that caused by statistical error [Davison and Hinkley 1997]. In the *integrated resampling* workflow we present a crowd member with a unique sequence of visualizations that depict different resamples drawn from the data. We show that this method, where the crowd member integrates multiple views to form a final assessment, results in order effects that systematically bias final assessments. In the *distributed resampling* workflow a crowd member is presented with a *single* unique visualization that is generated from the bootstrapped data. We show that aggregating distributed assessments do not suffer from order effects and generate additional distributional data the other methods do not.

We perform two experiments using Amazons Mechanical Turk to test whether order effects impact assessment in the *integrated resampling* workflow and to show how the *distributed resampling* workflow avoids order effects and produces other valuable information not obtainable from other approaches. In both experiments, a crowd member makes an assessment using visualizations generated using resampled datasets (bootstraps) drawn from a single data set.

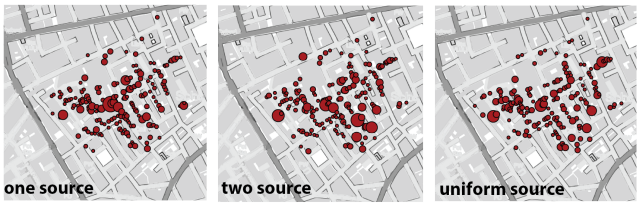

Fig. 1: Experimental stimuli suggesting “one source”, “two source”, and a “uniform distribution” of deaths. Red dots indicate deaths at that location, where the size of the dot represents the number of deaths.

To create the visualization stimuli used in our experiments, we bootstrapped a georeferenced version of John Snows cholera deaths dataset [Wilson 2007] that depicts death counts in the Soho district of London in 1854. Though Snow ultimately identified a single location (the Broad Street Pump) to be the source of the contamination, visual inspection of the whole data set does not clearly or intuitively suggest this conclusion [Koch 2005]. We recreate the task of visually assessing the source of the deaths but experimentally vary whether the resample maps appear to support a conclusion that deaths stem from a “single source”, from “two sources” or that there was no single source and that they appear “uniformly distributed”. To select our three visualizations, we create many resamples from the cholera deaths data and select maps that correspond most closely to the three opposing conclusions of “single source”, “two source”, and “uniformly distributed” deaths (by measuring death-to-pump distances and statistically testing the reference distributions which have a single, double, or no peaks). The final three visualization stimuli are depicted in Figure 1.

2. STUDY 1: TESTING FOR ORDER BIAS IN THE INTEGRATED RESAMPLING TASK

Experimental Procedure: In the *integrated resampling* task, a crowd member is shown two visualizations (one per screen) and is told that each is independent depiction, collected by researchers A and B, of deaths believed to be from the same disease within an urban area. After viewing the visualization from researcher A (on screen one), she is asked for a ’yes’/’no’ assessment of whether ’the deaths appear to be spreading from a single source. She also estimates the number of sources the deaths appear to be spreading from–ranging from 1 to 11,’ the maximum number of pumps in the Snow data–or can indicate that it appears Random with no identifiable source. After seeing the second visualization (from researcher B) on the second screen, she is prompted with: *Taking into account both researcher A and researcher B’s maps, please tell us whether you think that overall the deaths appear to be spreading from a single source?* At the top of the screen in bold type she is reminded of her initial assessment on the previous screen and of the number of sources she estimated.

The variable of interest is the final judgment and we specifically seek to answer: *does the order of the images influence the final assessment?*. The images are randomized so that crowd members see one of six possible orderings of two of the three images. Though not reported here, to confirm the robustness of the order effects we tested several versions of the visualization with different graphical features (e.g., blue dots to represent the pumps, labeling them as pumps, etc.) and found no effect for these variants. We present the results of the blue dots (no label) here.

Results: We test whether the order of visual evidence affects individuals overall judgment by examining the final judgments for those participants who saw the same two visualizations but in a different order. We have 161 participants after removing

	Indicator variables for order		
	OneTwo	**Two**Uniform	**One**Uniform
Order Indicator	-0.379 (0.083)***	0.280 (0.107)***	-0.041 (0.135)
Observations	62	55	54
Pseudo R^2	0.135	0.0745	0.0013

Notes: Probit: standard errors are in parentheses.
Significant: * 10 %, ** 5 %, *** 1 % level.

Table I: Marginal effects from probit regression of "overall judgment" responses on graph stimuli order. Columns contain results restricted to participants who saw the same two stimuli but in two different orders. The regressor "Order Indicator" equals 1 when they saw the first, bolded visualization first.

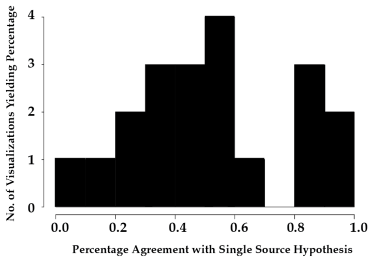

Fig. 2: A histogram showing the number of resample visualizations (out of 20) for different levels of agreement with the single source hypothesis.

inconsistent responses (e.g., an answer of "Yes" to the one source question with a selection of more than one source in the dropdown, or vice versa).

Table 1 reports the marginal effects from a probit regressing the overall judgment on visualization order (the dependent variable is coded as 1 when a crowd member's overall judgment was "Yes, single source" and 0 otherwise). Each column of Table 1 is a different regression testing the effect of order on a visualization pair. For example, column (1) tests whether the 62 crowd members who saw the "one source" followed by the "two source" visualization gave a different overall judgment than those who saw the "two source" followed by the "one source" visualization. Table 1 indicates that the presentation order of the same visualizations significantly affects crowd members overall judgments. When crowd members' first visualization is "one source" and their second visualization depicts "two sources", then they are 38% less likely (p <0.01) to make a final judgment of "single source" than when the order of presentation is reversed. When participants see a visualization that depicts "two sources" and then see one that depicts "uniform" sources, they are 28% (p <0.01) more likely to respond single source than if the order of the same visualizations is reversed. We see no significant order effect for the "OneUniform" condition. However, these results indicate that the order in which several visualizations appear can cause crowd members to unduly weight some pieces of evidence in their final judgments about data. Consequently, any crowd analysis workflow that requires crowd workers to integrate evidence across multiple visualizations can produce biased individual judgments that consequently add noise to the aggregated judgment.

3. STUDY 2: TESTING FOR ORDER BIAS IN THE DISTRIBUTED RESAMPLING TASK

In study 2, we show that the *distributed resampling* approach avoids the bias we observe in study 1 and compare results from the distributed resampling approach to those from a *one-to-many* procedure.

Experimental Procedure: In the *distributed resample* treatment we randomly select 20 visualizations, generated by bootstrapping the Snow maps, and assign them to 200 crowd members (each image is assessed by approximately 10 individuals) We compare overall judgments in the distributed resample treatment to the overall judgments made by crowd members who saw a single visualization. In our *one-to-many* baseline we present an additional 150 crowd members with a single visualization of Snows original data. In the baseline and treatment, the task is described to crowd members just as in Study 1, except that they know only one visualization will be shown.

Results: We have 260 total responses (one-to-many: 117, distributed: 143) after removing inconsistent responses. For the one-to-many visualization, the rate of responding single source was 59.0% (69/117). We find that 45.4% of crowd members state that is was a single source when we aggregate the responses obtained from the distributed cognition approach. We use Wilsons score interval calculation (Wilson 1927) for binominal percentage confidence intervals to construct intervals from 49.9% to 67.4% for the one-to-many treatment and from 37.5% to 53.6% for the distributed resampling treatment. A two sample proportion test indicates a significant difference between the distributed and one-to-many percentages (X^2=4.18, df=1, p <0.05).

Crowd members were more likely to state that there was one source in the one-to-many treatment. The interval obtained by using the distributed approach is similar in size but centered at a lower level of agreement with the single source hypothesis. In this case, the observed rate of agreement with the single source hypothesis is obtained through a balance of diversity (obtained by randomly drawing resampled visualizations) and replication (obtained by showing each selected resampled visualization to multiple analysts). The aggregated level of agreement with the single source hypothesis is therefore more likely to reflect a sensitivity to variance in the data. The researcher can examine the distribution of responses across the set of samples (Fig. 2) to further assess the certainty of the aggregate judgment. Even if crowd members participating in the one-to-many approach were cautious about the potential for uncertainty, the approach does not support similar estimation of the certainty of the crowd judgment.

4. CONCLUSIONS

We propose a resampling and visualization procedure and distinguish two viable crowdsourced workflows: an integrated cognition resampling approach and a distributed cognition resampling approach. Our experimental results show that analysts are biased by viewing order when integrating information. The distributed cognition resampling approach avoids order effects by allocating a single visualization to each crowd member. By comparing the distributed approach to a "one-visualization-to-many" approach, we demonstrate another benefit of the distributed approach: distributional information that allows better estimation of a margin of error.

REFERENCES

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4 (Dec. 2005), 389–396. DOI:http://dx.doi.org/10.1037/1082-989x.10.4.389

A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and their Application* (1 ed.). Cambridge University Press.

Jeffrey Heer. 2007. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In *Proceedings of the CHI07*. 1029–1038.

Tom Koch. 2005. *Cartographies of Disease: Maps, Mapping, and Medicine* (1st ed.). ESRI Press.

Jill H. Larkin and Herbert A. Simon. 1987. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11, 1 (1987), 65–100. DOI:http://dx.doi.org/10.1111/j.1551-6708.1987.tb00863.x

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

Martin Wattenberg and Jesse Kriss. 2006. Designing for social data analysis. *IEEE transactions on visualization and computer graphics* 12, 4 (Aug. 2006), 549–557. DOI:http://dx.doi.org/10.1109/TVCG.2006.65

Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 227–236. DOI:http://dx.doi.org/10.1145/2207676.2207709

Robin Wilson. 2007. John Snow's famous cholera data analysis data in modern GIS formats. (Jan. 2007). http://blog.rtwilson.com/john-snows-famous-cholera-analysis-data-in-modern-gis-formats/