Unbiasing Visual Data Exploration in the Garden of Forking Paths

Xiaoying Pu
xpu@umich.edu
Computer Science and Engineering,
University of Michigan
Ann Arbor, MI

Matthew Kay
mjskay@umich.edu
School of Information, University of Michigan
Ann Arbor, MI

Michael Correll
mcorrell@tableau.com
Tableau Research, Tableau Software
Seattle, WA

Eli T. Brown
eli.t.brown@depaul.edu
College of Computing and Digital Media,
DePaul University
Chicago, IL

KEYWORDS

Bias, exploratory data analysis, data wrangling, uncertainty visualization

ACM Reference Format:
Xiaoying Pu, Matthew Kay, Michael Correll, and Eli T. Brown. 2019. Unbiasing Visual Data Exploration in the Garden of Forking Paths. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

INTRODUCTION

We are interested in a kind of *bias* in the process of *data exploration*. In the current exploratory data analysis (EDA) tools, there is a danger of wandering in the *garden of forking paths* [5]: analysts trying alternative paths to explore data and taking "interesting" data patterns as confirmatory, leading to biased and non-generalizable conclusions. This kind of bias can take the forms of multiple-comparison problem, overfitting, and more. In previous work, we describe how all of these biases can be considered

Conference'17, July 2017, Washington, DC, USA
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Author Background

Xiaoying Pu is a Ph.D. student in Computer Science and Engineering at the University of Michigan. She designs and evaluates visualizations to make data exploration more statistically reliable. Her website is xiaoyingpu.github.io.

Matthew Kay is an Assistant Professor in the University of Michigan School of Information working in human-computer interaction and information visualization. He studies the communication of uncertainty in domains like personal informatics, everyday sensing and prediction, and scientific communication. He has published work advancing the use of Bayesian statistics in VIS and CHI. His website is: www.mjskay.com.

Author Background

Michael Correll is a research scientist at Tableau Software working in the area of information visualization and statistical communication. He has published work on uncertainty visualization in VIS and CHI. His website is correll.io.

Eli T. Brown is an Assistant Professor in the DePaul University College of Computing and Digital Media. His research integrates visual analytics and machine learning in problem domains including journalism, biotechnology, and medical informatics, where decision making under uncertainty is necessary. He has published work on machine learning from user interactions in VIS and CHI. His website is: elitbrown.org.

[1]Null hypothesis significance testing

to be the *forking paths problem*: unaddressed flexibility in data analysis that leads to unreliable conclusions [14]. The forking paths problem is significant and real: it can be considered as one of the causes of the replication crisis, and previous participant experiments have caught it in action [17].

Our idea and on-going work is to make data exploration more robust to the forking paths problem in visual analytics. Since visualizations are important exploration tools, we wish to design and evaluate visual representations that correct for this problem. We want to first identify and then deploy effective visualization techniques to represent the inherent uncertainty in a dataset. Novel representations backed by statistical techniques may encourage healthy skepticism about exploratory "insights" from visualizations. Ultimately, we aim to improve data exploration outcomes, i.e., analysts making more optimal decisions about "interesting" data patterns.

The forking paths problem: biases from data exploration

The current EDA systems can be dangerous in two ways: 1) they do not surface or adjust for intrinsic data uncertainty such as sampling error, and 2) they do not explicitly separate exploratory and confirmatory analyses (no validation). Taking exploratory findings as confirmatory is in theory "destructively foolish" [15]. In practice, analysts might not clearly distinguish exploratory and confirmatory findings. Based on an interview study, Alspaugh *et al.* report that even professional analysts would do *ad hoc* exploration even though the analysts are wary of the practice [1].

The implications of the forking paths problem can be profound: it can result in analysis biases, such as overfitting in model-fitting and multiple comparison problems in NHST[1]. "p-hacking" is a special instance of the forking paths problem where researchers hunt for and publish significant p values only and consequently, multiple important findings in social psychology and beyond failed to be replicated [12]. Since we believe that the forking paths problem does not typically arise from malicious user intentions, we envision system designs that steer analysts away from the forking paths problem.

We focus on visual analytics tools, where analysts can select, filter, and zoom into interesting patterns in the data, e.g. TimeSearcher [7]. However, the convenience of visual analytics can also exacerbate the forking paths problem. Gotz *et al.* call visualizations "visual predictive models", meaning that the surface value of visualizations are commonly taken as basis for decision-making [6] while in reality, the visualizations do not portray data uncertainty. As evidence to the forking paths problem, Zgraggen *et al.* have shown that when an analyst operates in the NHST framework, unchecked exploration in visual analytics can lead to higher false discovery rates (FDR) in an experimental setting [17]. The forking paths problem may also be exacerbated by automatic insight discovery and recommendation systems, such as SeeDB [16]. These systems might further encourage and enable analysts to pay attention to interesting but potentially false patterns, which Correll elevates to an ethical concern [4].

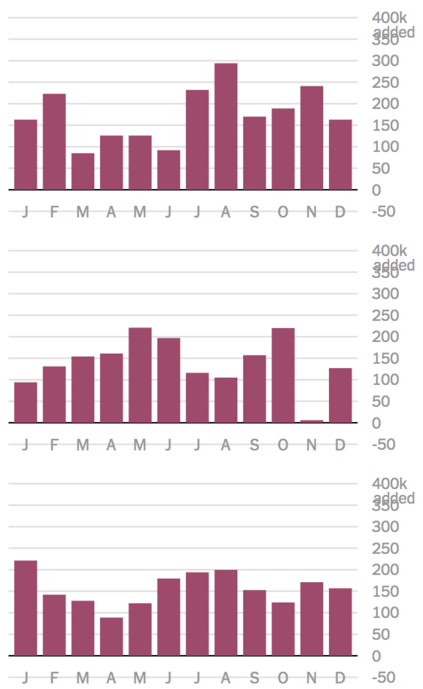

Figure 1: Some frames in a hypothetical outcome plot. These figures describe what "jobs added (in 1,000) over twelve months" could be like, given the model that assumes job growth is actually constant. Screenshots taken from New York Times [10].

OUR IDEA: UNBIASING VISUAL ANALYTICS

Given the significance of the forking paths problem, we want EDA systems to convey the bias in sample data, so that exploratory findings are not easily taken as confirmatory. One proposed way to curtail the forking paths problem is to use a public blockchain to track all significance tests performed on a dataset [3]. However, we want to take a more human-centered approach and accommodate more than NHST methods. Since visualizations are important tools in the data wrangling process [11], we imagine alternative ways to visualize data in an EDA system such that the forking paths problem is transparent to the analyst, encouraging healthy skepticism towards potentially biased insights. Overall, we want a visual analytics system to be "truthful" [2] in enabling insights and decisions.

More concretely, we propose to infuse several statistical techniques into visualizations. We may be able to discourage analytical biases while hiding technical details such as p-values and false discovery rates from the users. Our previous work has laid out a design space for visual analytics to address the forking paths problem, where we disentangle visual/data representation options (annotations and data transformations) and statistical techniques (regularization and multiple-comparison correction) [14]. Extending our design space, we also want to incorporate modern uncertainty visualizations and the bootstrapping statistical technique.

One such modern uncertainty visualizations is *hypothetical outcome plots* (HOPs) [9]. Instead of a traditional static visualization, a HOP is an animation where each frame shows one possible data sample from an underlying distribution, see Figure 1. Uncertainty visualizations such as HOPs have the same, familiar visual encoding as the traditional visualizations, allowing the system to easily communicate data uncertainty that arises from the forking paths problem. We can generate each frame of a HOP with *bootstrapping*. Bootstrapping is a generalizable technique where the original dataset is sampled with replacement, creating alternatives of "what the data could have looked like".

We are presently building and evaluating an EDA system that employs techniques described above to mitigate the forking paths problem, as we describe in our previous paper [14]. As recommended by Hullman *et al.* [8], we plan to use incentivized decision-making experiments to evaluate different techniques to address the forking paths problem.

MOTIVATIONS FOR PARTICIPATING IN THIS WORKSHOP

Through the workshop, the authors would like to bring more awareness to the forking paths problem in data exploration and get feedback on the planned evaluations. Given the broader theme of this workshop, they are also interested in discussing how knowledge about data science work topics can inform experimental and design work:

- What are analysts' incentives to find "interesting" patterns?
- How do analysts make decisions about what to report?

- How are biased, exploratory insights formed, and how do they propagate through the analysis pipeline, and ultimately effect real-life decision-making outcomes?
- How can we tell if an analyst's *intention* [13] is to conduct exploratory or confirmatory analyses?

REFERENCES

[1] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A Hearst. 2018. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. (2018), 10.

[2] Alberto Cairo. 2016. *The Truthful Art: Data, Charts, and Maps for Communication* (1st ed.). New Riders Publishing, Thousand Oaks, CA, USA.

[3] Yeounoh Chung, Sacha Servan-Schreiber, Emanuel Zgraggen, and Tim Kraska. 2018. Towards Quantifying Uncertainty in Data Analysis & Exploration. (2018), 13.

[4] Michael Correll. 2018. Ethical Dimensions of Visualization Research. *arXiv:1811.07271 [cs]* (Nov. 2018). arXiv:cs/1811.07271

[5] Andrew Gelman and Eric Loken. 2013. The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Time. *Department of Statistics, Columbia University* (2013). https://doi.org/dx.doi.org/10.1037/a0037714

[6] David Gotz, Wenyuan Wang, Annie T Chen, and David Borland. 2019. Visualization Model Validation via Inline Replication. *Information Visualization* (Jan. 2019), 1473871618821747. https://doi.org/10.1177/1473871618821747

[7] Harry Hochheiser and Ben Shneiderman. 2004. Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration. *Information Visualization* 3, 1 (March 2004), 1–18. https://doi.org/10.1057/palgrave.ivs.9500061

[8] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2019), 903–913.

[9] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[10] Neil Irwin and Kevin Quealy. 2018. How Not to Be Misled by the Jobs Report. *The New York Times* (Jan. 2018).

[11] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data. *Information Visualization* 10, 4 (Oct. 2011), 271–288. https://doi.org/10.1177/1473871611415994

[12] Open Science Foundation. 2015. Estimating the Reproducibility of Psychological Science. *Science (New York, N.Y.)* 349, 6251 (2015), aac4716. https://doi.org/10.1126/science.aac4716 arXiv:1011.1669v3

[13] William A. Pike, John Stasko, Remco Chang, and Theresa A. O'Connell. 2009. The Science of Interaction. *Information Visualization* 8, 4 (Jan. 2009), 263–274. https://doi.org/10.1057/ivs.2009.22

[14] Xiaoying Pu and Matthew Kay. 2018. The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics. 9.

[15] John W Tukey. 1977. *Exploratory Data Analysis.* Addison-Wesley Pub. Co., Reading, Mass.

[16] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. SeeDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *Proc. VLDB Endow.* 8, 13 (Sept. 2015), 2182–2193. https://doi.org/10.14778/2831360.2831371

[17] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, New York, New York, USA, 1–12. https://doi.org/10.1145/3173574.3174053