Toward A Better Understanding of Missingness in Visual Analytics

Yue Ma* Courtney Bolton* Maoyuan Sun* Yuanxin Wang† Jian Zhao† Tianyi Li‡
*Northern Illinois University †University of Waterloo ‡Purdue University

ABSTRACT

Data-driven decision making has been a common task in today's big data era, from simple choices such as finding a fast way for driving to work, to complex decisions on cancer treatment in healthcare, often supported by visual analytics. For various reasons (e.g., an ill-defined problem space, network failures or bias), visual analytics for sensemaking of data involves missingness (e.g., missing data and incomplete analysis), which can impact human decisions. To be aware of missingness, as an initial step, we present a framework of categorizing missingness in visual analytics from two perspectives: *data-centric* and *human-centric*. The former emphasizes missingness in three data-related categories: *data composition*, *data relationship* and *data usage*. The latter focuses on the human-perceived missingness at three levels: *observed* missingness, *inferred* missingness and *ignored* missingness. Based on the framework, we discuss possible roles of visualizations for handling missingness, and conclude our discussion with future research opportunities.

Index Terms: Human-centered computing—Human Computer Interaction; Human-centered computing—Visualization

1 INTRODUCTION

For various reasons (e.g., network problems, system design failures or bias), *missingness* often exists in a human sensemaking process. It impacts human decisions and can lead to severe consequences. In fact, a successful awareness of missingness is an initial but critical step for handling missingness, and visualization can help with missingness awareness for data analytics [2, 3]. In this work, we investigate missingness from a sensemaking perspective. The goal is to establish a systematic understanding of missingness and pave the road for future research using visual analytics to handle issues related to or caused by missingness.

As an initial step, we present a framework to categorize missingness that may exist in a sensemaking process with visual analytics. This framework considers missingness from two perspectives: *data* and *human*. The two perspectives consider the two key parties in visual analytics: computation and human cognition, which are glued together by interactive visualizations. This framework enables a systematic consideration of missingness in visual analytics. Relying on it, to handle missingness, the visualization design needs to reveal possible data-related missingness and aims to prevent users from simply ignoring missingness. We hope this work can draw attention to future exploration of the design space of visualizing missingness and studying insights from missingness in sensemaking.

2 A DATA-CENTRIC VIEW OF MISSINGNESS

A data-centric view categorizes missingness in three data-related groups: *data composition*, *data relationship* and *data usage*.

*e-mail: Z1934458@students.niu.edu, {cbolton1, smaoyuan}@niu.edu.
†e-mail: {y2587wang, jianzhao}@uwaterloo.ca.
‡e-mail: li4251@purdue.edu

2.1 Missingness in Data Composition

There are three possible types of missingness, listed as follows:

Missing data entities highlights the absence of data entities. It often comes from errors in the process of data collection [1]. For example, in fitness tracking devices, some sensor data might not be successfully recorded due to network connection failures.

Missing data attributes reveals the incompleteness of data attributes. This may come from a careless design of the data collection mechanism [7]. For example, when creating a survey, researchers might fail to include all the relevant.

Missing data values (often named as missing data) is the loss of data values. Compared to the other two, it has drawn the most attention and been heavily studied [4, 5, 10].

2.2 Missingness in Data Relationship

Missing data relationships refers to the absence of relations among data entities. From a graph perspective, it highlights the lack of links among nodes in a graph. This means that for a given set of data entities, some connections between data entities are not present. An absence of relationships among data entities may either result from errors in a data collection process or be a reflection of algorithmic results of data relationship discovery.

2.3 Missingness in Data Usage

The utility of data for sensemaking activities involve two key types: 1) *data selection* and 2) *analytical method selection*. The former refers to which parts of a given dataset will be selected for analysis. The latter means which analytical methods will be picked and applied to the selected data. Missingness in data usage can happen in both activities due to uncertainties and selection biases [6, 11]).

3 A HUMAN-CENTRIC VIEW OF MISSINGNESS

A human-centric view categories missingness at three levels: 1) *observed* missingness, 2) *inferred* missingness and 3) *ignored* missingness. They reveal how the data-centric missingness is perceived by people.

3.1 Observed Missingness

Observed missingness means that users can directly perceive missingness. It indicates that the visibility of missingness is high, and users can easily notice it. As the visibility of missingness is affected by the way that data is represented, observed missingness relies on the visual context, in which data is encoded by certain visualizations. Different visual encodings can impact how easily users can observe missingness. For example, it is easier for users to see missing links by looking at a matrix than checking the same data displayed lists of node-pairs. Thus, for observed missingness, users can verify their perceived missingness by referring to the given visual context (e.g., pointing to an empty cell).

3.2 Inferred Missingness

Inferred missingness refers to that the visibility of missingness goes low or missingness even gets invisible, so it is impossible for users to directly observe missingness. However, via an investigation with given data, users can infer the possible existence of missingness. Compared to observed missingness, inferred missingness may not

be easily verified. Thus, observed missingness is more confirmative, while inferred missingness is more hypothetical.

3.3 Ignored Missingness

Ignored missingness indicates no observation nor awareness of missingness, or the presence of possible missingness is not considered. It may appear for two reasons. First, the visibility of missingness is too low to raise user awareness. For example, a user may never realize that missing edges exist after looking at lists of edges. Second, due to some biases or the impact of cognitive capture (or tunneling) [9], users turn a blind eye to possible missingness. For example, to explore possible treatment for a disease, all effort has been put on the group of people who have been infected by the disease, while the uninfected group never gets any attention.

4 Handling Missingness: The Role of Visualization

Based on the data-centric and human-centric perspectives of missingness mentioned before, in this section, we discuss four possible roles of visualizations for supporting missingness handling. The first and second roles highlight supporting the detection of data-centric missingness. The other two roles aim to improve user awareness of data-related missingness.

Bridging Existing Data and Missing Data: Visualizations play a key role of bridging the gap between existing data and missing data. To establish such a bridge, a commonly used strategy is *space-filling* that reveals missingness as empty (e.g., an empty space in a bar chart [10]), gap (e.g., broken lines in a line chart [10]), or different-looking space (e.g., a matrix with different colored cells [4, 12]). By looking at the visually salient space, users can be aware of data-centric missingness.

Supporting the Analysis of Analytic Provenance: To help avoid data usage related missingness, visualizations can be used to support tracking analytical provenance and further analyze it. Two key aspects need to be considered in this process: 1) the selected, investigated, derived and newly generated data, and 2) the method or process applied to such data. They, respectively, correspond to the provenance of data and process [8]. Visualizing them offers a way of analyzing analytic provenance.

Improving Awareness: from Ignoring to Observing: From a perceptual-oriented perspective, a key role of visualization is to prevent users from falling in the trap of ignoring data-centric missingness. This implies that using visualizations can improve the *expressiveness* of data-centric missingness. The higher such expressiveness goes, the easier it is for users to observe possible missingness.

Scaffolding Missingness Inference: Visualizations offer a usable mean to scaffold missingness inference. In this case, visualizations may focus on displaying either the connections across different parts of data or the provenance of a sensemaking process. These help users to infer possible existence of data-centric missingness. Since inference is a reasoning process, instead of a static stage, it has more complex needs for the design of visualizations, multiple types of visualizations may be used and fusing information across them can be helpful to scaffold missingness inference.

In summary, visualizations can assist to uncover the data-centric missingness and improve their expressiveness, so they become more visible and accessible to users.

5 Discussion and Conclusion

While handling missingness remains a challenging problem in sensemaking, as an initial exploration, we present a framework that helps to systematically view and categorize missingness in visual analytics. It highlights considering missingness from two key perspectives: *data-centric* and *human-centric*. The former regards missingness in three data-related categories: *data composition*, *data relationship* and *data usage*. The latter focuses on the human-perceived missingness at three levels: *observed* missingness, *inferred* missingness and *ignored* missingness. Based on the framework, we discuss four possible roles of visualizations for helping to handle missingness in a sensemaking process, and there are three research themes that are worthy of future studies: 1) missingness detection, 2) missingness visualization and 3) missingness insight.

Detecting Missingness: Missingness detection lays the foundation for effective data analysis. If users were not clear about which types of data-centric missingness exist, it would be hard for them to further explore and work on detection methods. Also, a sensemaking process can have multiple types of data-centric missingness. Our framework may help to clarify the detection goals.

Visualizing Missingness: The design of missingness-oriented visualizations remains an under-explored direction. How to formalize the design space of missingness visualizations? How and if possible can we measure the expressiveness of visual encodings for data-centric missingness? The perceptual-perspective discussed in our framework may help to derive usable measures.

Discovering insights from Missingness: Studying possible insights that users gain from missingness in sensemaking is a highly sought-after research challenge. The insights derived from missingness may depend on an application domain and different types of data-centric missingness may bring different insights. It may further broaden our understanding of evaluating visualizations by considering the value of missingness.

In summary, we present a framework that provides a systematic view of missingness in visual analytics. We hope this work can draw attention to future studies on visual sensemaking with missingness.

References

[1] P. D. Allison. *Missing data*. Sage publications, 2001.

[2] R. Andreasson and M. Riveiro. Effects of visualizing missing data: an empirical evaluation. In *International Conference on Information Visualisation*, pp. 132–138. IEEE, 2014. doi: 10.1109/IV.2014.77

[3] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. In *IFIP Conference on Human-Computer Interaction*, pp. 861–872. Springer, 2005. doi: 10.1007/11555261_68

[4] S. J. Fernstad. To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. *Information Visualization*, 18(2):230–250, 2019. doi: 10.1177/1473871618785387

[5] S. J. Fernstad and R. C. Glen. Visual analysis of missing data—to see what isn't there. *2014 IEEE Conference on Visual Analytics Science and Technology*, pp. 249–250, 2014. doi: 10.1109/VAST.2014.7042514

[6] J. Heckman. Varieties of selection bias. *The American Economic Review*, 80(2):313–318, 1990.

[7] T. D. Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383, 2001. doi: 10.1076/edre.7.4.353.8937

[8] Y. L. Simmhan, B. Plale, D. Gannon, and S. Marru. Performance evaluation of the karma provenance framework for scientific workflows. In *International Provenance and Annotation Workshop*, pp. 222–236. Springer, 2006. doi: 10.1007/11890850_23

[9] D. J. Simons and C. F. Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, 28(9):1059–1074, 1999. doi: 10.1068/p281059

[10] H. Song and D. A. Szafir. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics*, 25(1):914–924, 2018. doi: 10.1109/TVCG.2018.2864914

[11] E. Wall, J. Stasko, and A. Endert. Toward a design space for mitigating cognitive bias in vis. In *2019 IEEE Visualization Conference (VIS)*, pp. 111–115. IEEE, 2019. doi: 10.1109/VISUAL.2019.8933611

[12] J. Zhao, M. Sun, F. Chen, and P. Chiu. Missbin: Visual analysis of missing links in bipartite networks. In *2019 IEEE Visualization Conference (VIS)*, pp. 71–75. IEEE, 2019. doi: 10.1109/VISUAL.2019.8933639