Ecological Validity in Quantitative User Studies – a Case Study in Graph Evaluation

Mershack Okoe* Radu Jianu†

Florida International University

ABSTRACT

Quantitative user studies are often reviewed and judged by the magnitude of detected effects and basic soundness of their evaluation procedure. Here, we advocate for an increased focus on the ecological validity of tasks, interactions, and data chosen for evaluation. We revisit Ghoniem et al.'s highly cited study of node-link vs. matrix representations of graph data [2], discuss the ecological validity of its design using a formal framework, and show quantitatively how minor changes in task and interaction phrasing can lead to significantly different outcomes and insights.

Keywords: User studies, node link diagrams, adjacency matrices.

1 INTRODUCTION

We argue that controlled user studies, which are central to data visualization research [1], are too often judged only in terms of effects detected and basic soundness of their protocol, in detriment of their ecological validity. We show how considering ecological validity of a study's tasks, interactions, and data, can lead to important differences in evaluation outcomes and conclusions. Specifically, we revisit the highly cited study by Ghoniem et al. [2] which compared node-link diagrams (NLD) and adjacency matrices (AM), and found that for large graphs, AM performed better than the NLD in both accuracy and time for all of seven tasks. We discuss the ecological validity of the study within a formal framework, then show quantitatively that testing the same fundamental 'data-reading' tasks but with slightly modified tasks and interactions can lead to different conclusions.

2 ORIGINAL STUDY

Ghoniem et al. [2] compared NLDs and AMs on seven graph tasks. Users had to: 1-estimate node count, 2-estimate edge count, 3-find the most connected node, 4-find a node by its label. Given two selected nodes they had to find: 5-if they are connected, 6-if they share a neighbor, and 7-if there is a path between them. Users could select multiple nodes and highlight another via mouse-over in both representations. Randomly generated graphs of three sizes (20, 50, 100) and densities (0.2, 0.4, and 0.6) were used. The AM was sorted lexicographically.

Results: For small sparse graphs, NLD and AM were similar, but NLD was better in connectivity tasks (5, 6, 7). For large and dense graphs, AM outperformed NLD in all seven tasks.

3 DISCUSSION

We formalize our discussion of ecological validity into a framework of five questions, which may be generalizable to evaluations beyond the current case study.

*e-mail: mokoe001@cis.fiu.edu

†e-mail:rdjianu@cis.fiu.edu

Q1: Is the study using ecologically valid data? Real-life graph data rarely exhibits random topological structure. Moreover, an important benefit of graph visualizations is that they can reveal such structure. As such, random graphs may not be a particularly ecologically valid choice of data.

Q2: Is the presentation of the visualizations ecologically valid? Ghoniem et al. used lexicographically ordered AMs. These support the tasks they evaluated well. For example, it is unsurprising that finding a node takes constant time in their AMs, as this task reduces to scanning an ordered list of labels. However, lexicographic AMs do not reveal important topological properties, and may be used less often than those that can (Figure 1).

Q3: Are the visualizations equivalent? We argue that NLDs are not equivalent to lexicographic AMs, since the first reveals structure while the second cannot. While it is true that the visualizations are equivalent for the subset of evaluated tasks, a complete answer needs to consider (i) how often are lexicographical AMs used, especially if topological ordering is also available, and (ii) how often do users change AM ordering depending on their tasks. We believe the use of AM that expose topological structure (Figure 1) would have led to a more meaningful comparison.

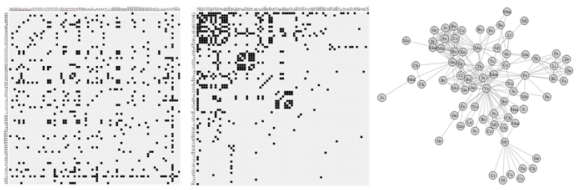

Figure 1: Lexicographically ordered AMs (left) cannot reveal graph structure in the same way that a clustered AM (center) and an NLD (right) can.

Q4: Are the interactions equivalent? This is a difficult question because: (i) an interaction in one visualization may not have an equivalence in another; (ii) the same interaction may aid each visualization in different ways and to different degrees. For example, Ghoniem et al.'s selection of nodes in the AM is not equivalent to the one in the NLD. As shown in Figure 2, the NLD allows us to easily read the neighbors of a selected node since they are exposed by their incident edges. This is more difficult in the matrix since a user has to trace from a dot vertically or horizontally through the matrix to reach its label, without any visual aid. As we will show, this difference becomes important if the connectivity task is phrased differently.

Furthermore, an interaction that most NLDs implement is that of moving a picked node. This interaction can often clarify where a selected node's edges end in a dense visualization. It does not however have an equivalence in the AM, and Ghoniem et al. have not added it to their NLDs. As we will show, the absence of this feature was the main reason of poor performances by Ghoniem et al.'s users when using NLDs on large graph connectivity tasks.

This begs an important question: does adding this feature give an unfair advantage to NLD? We think not, because any interaction

involves a cost in addition to a benefit. As long as the interaction is useful and part of how the visualization is typically used, it is ecologically valid and should not be abstracted away.

Q5: Are the chosen tasks, as presented to subjects, ecologically valid? As defined, this question has two components: (i) is the fundamental tasks valid, and (ii) is the task phrasing valid?

For example, most would agree that determining if two nodes are connected is a fundamental graph task. However, this task presents in many forms: a user may look at a highlighted or unhighlighted node to read its neighbors, or at two highlighted or unhighligted nodes to determine if they share an edge. These four scenarios are not equivalent as any interaction involves an overhead. Ghoniem et al. evaluated the connectivity task by highlighting both nodes which we argue may be the least ecological instantiation of this fundamental task: exploring a graph does not generally rely on pairwise node selections.

More generally, two questions can help quantify the ecological validity of a task: (i) how often do real users perform the task as phrased in the study?; (ii) can a task be easily replaced by an equivalent, more efficient interaction or query? While the first question is somewhat evident, the second bears discussion. Ghoniem et al.'s first three tasks could easily be implemented as graph queries, in the same way text editors offer functionality for counting words. Locating nodes by quering is also available in most visualization systems, and finding a node should take constant time once the cost of visual search exceeds that of typing. More broadly, if a visual task can be replaced by a query that can be posed and computed faster, then the visual task may have limited ecological validity.

4 User Study

Hypotheses: We hypothesized that a user study following the aforementioned guidelines would yield different conclusions than those of Ghoniem et al. Specifically, we focused on two scenarios in which AM outperformed NLD: task 5 ('connectivity task') and task 6 ('common neighbor task'), both for large graphs. We made the following changes to Ghoniem et al.'s study: (i) we used a real data set; (ii) we ordered the AM to reveal topological structure; (iii) we allowed users to drag nodes in NLDs; (iv) we created two versions of task 5: one using the original phrasing, in which both nodes are selected (5a), and a new task in which one node is selected while the other is named by its label (5b). Our hypotheses were that given these changes:

H1: NLD will outperform AM for both tasks 5a and 5b, even for large graphs. Reason: moving nodes allows NLD users to better see where edges end.

H2: AM will perform worse on 5b than 5a. Reason: the AM selection, as implemented, is less powerful than the NLD one.

H3: NLD will outperform AM for task 6, even for large graphs. Reason: moving nodes allows NLD users to better see where edges end.

Protocol and delivery: We designed a 2 visualizations x 3 tasks between-groups study, and used GraphUnit [3] to run it online via Amazon Mechanical Turk (MT). We drew the NLD using D3's generic forced directed method, and we ordered the AM using public D3 code. The underlying data was a graph of 100 nodes and link density 0.2, derived from a book recommendation dataset. We changed the book names to match the simplified nomenlacture used by Ghoniem et al. (i.e., A0..F9). We provided the same interactions as Ghoniem et al., namely node selection and node highlighting (Figure 2), and added node dragging in NLD. Formally, we evaluated the following tasks:

5a: Given two highlighted nodes, determine if they are connected.

5b: Given one highlighted node and the label of a second, determine if they are connected.

6: Given two highlighted nodes, determine if they have a common neighbor.

Following an introduction, subjects trained on five instances of each task type (15 training tasks), then completed the study with another five instances of each type (15 actual tasks). To minimize boredom and learning effects between the three evaluated tasks, we alternated the order in which we presented them to users. We recruited a total of 90 Mechanical Turk (MT) users, 45 for each of the two visualizations.

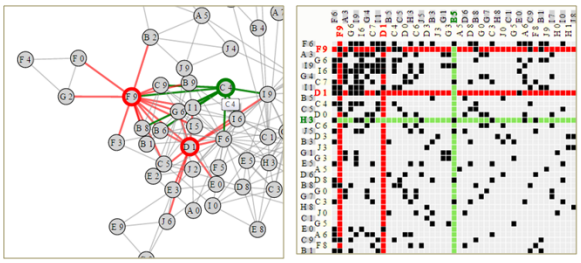

Figure 2: Available interactions: hovering/selecting a node highlights it and its edges green/red; hovering a link highligts it and its endpoints green. The images illustrate a task 5a instance.

5 Results

A Shapiro-Wilk analysis of our users' time and accuracy showed it was not normally distributed. We thus used a Wilcoxon-rank-sum test to analyze both time and accuracy. Our results were different from those of Ghoniem et al. We found that NLDs were more accurate than AMs for tasks 5a ($p < 0.001$), and both more accurate and faster ($p < 0.001$, $p = 0.002$) for 5b. This confirms H1 and H2. Finally, the NLD is significantly more accurate than AM for task 6 ($p < 0.001$), thus confirming hypothesis H3.

Our contributions are three-fold. First, we provide a framework for discussing the ecological validity of visualization user studies, and demonstrate its applicability in a case study. Second, we show how even small changes in study setup can lead to different outcomes. Third, we explain some of Ghoniem et al.'s surprising results (e.g., inability to move nodes determined the lower NLD performance), and end up with a different recommendation: NLDs are significantly better for all evaluated topological tasks, regardless of graph size and density.

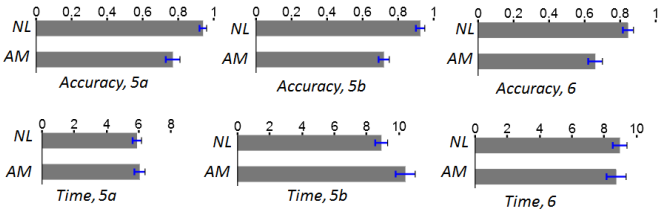

Figure 3: Accuracy and time results for the three tasks.

References

[1] S. Carpendale. Evaluating information visualizations. In *Information Visualization*, pages 19–45. Springer, 2008.

[2] M. Ghoniem, J. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 17–24. IEEE, 2004.

[3] M. Okoe and R. Jianu. Graphunit: Evaluating interactive graph visualizations using crowdsourcing. *Computer Graphics Forum (Proceedings of Eurovis 2015)*, 34(3), 2015.