Modeling the Complexity of Sankey Diagrams

Shashank Ginjpalli, Computer Science
Mentor: Dr. Chris Bryan, Assistant Professor
School of Computing, Informatics and Decision Systems Engineering

Can we successfully quantify the complexity of Sankey Diagrams?

Abstract

In this project, the team studied the complexity of Sankey diagrams, which are a type of visualization technique that shows flow between groups. To do this, the team created a carefully controlled dataset of synthetic Sankey diagrams of varying sizes as study stimuli. Then, a pair of online crowdsourced user studies were conducted and analyzed. User performance for Sankey diagrams of varying size and features (number of groups, number of timesteps, and number of flow crossings) were used to determine if they matched an initial formula.

Research Goals

- Develop a method to quantify the complexity of Sankey diagrams
- Evaluate this method by running a series of user studies to determine its performance against the visual complexity

Current Research Progress

- **Dataset Generation:** 3 sizes of Sankey Diagrams were generated using the Plotly Python library with controlled factors being the number of groups, number of timesteps, and the amount of starting flow. These are shown in Figure 1
- **Initial Complexity Formula:** A basic complexity formula was developed to model the complexity of the dataset. $complexity = timesteps + groups + flows + intersections$
- **Balance Distribution:** Initially plotted distribution of complexity in a histogram and noticed an over-representation of low complexity diagrams in the dataset and had to change some of small sized diagrams to medium sized
- **User studies:** Two user studies were conducted one measuring the performance of users and one measuring the visual complexity of the diagrams in the dataset
- **Statistical Analysis:** A basic analysis was performed measuring the average performance of users in the first study and performing a pair-wise ranking on the second study versus the complexity computed by the formula

Results & Analysis

Study 1: Sankey Summarization: Chart 1 shows the results of the first study. Each diagram in the dataset had 4 possible questions that could be asked about it. The chart plots the overall question accuracy versus the diagram complexity for each of the diagrams. As shown in the chart we observed that the question accuracy dropped when participants were faced with more complex diagrams

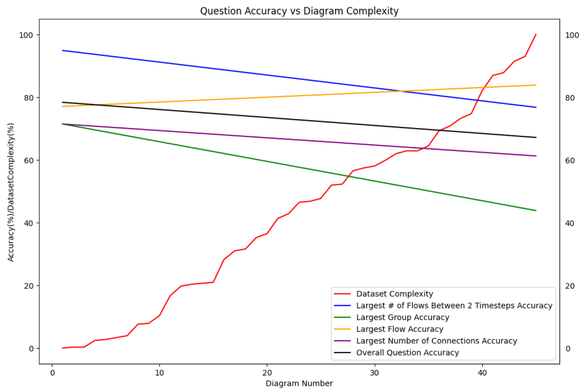

Chart 1: This chart shows the % of questions associated with each diagram that were answered correctly plotted against the complexity of each diagram.

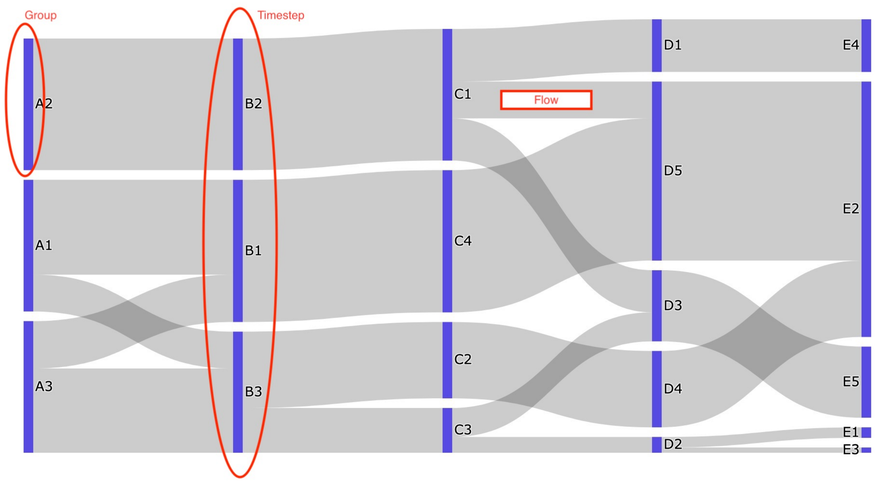

Figure 1: The parts of the Sankey Diagram that were controlled during the data generation along with providing an example of a Sankey Diagram

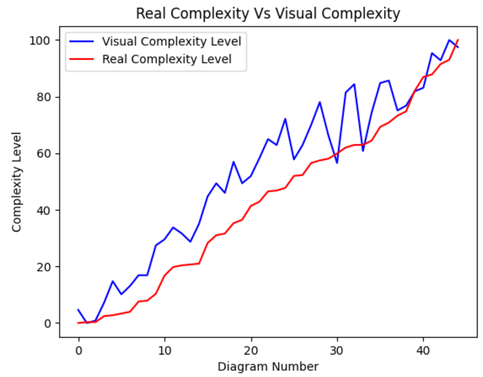

Chart 2: This chart shows the visual complexity from study 2(blue line) plotted against the real complexity. As shown by this figure the Visual Complexity can be closely modeled by the complexity formula line shown in red.

Study 2: Sankey Comparison: In this study participants were asked to compare 2 diagrams from the dataset. Pair-wise ranking was performed to determine if the visual complexity ordering of the diagrams matched up to the computed complexity ordering. These results were then normalized and plotted versus the visual complexity score as shown in Chart 2.

Future Work

The results of the 2 studies indicate that visual complexity can be modeled but the complexity formula is currently just an estimate

- **ANCOVA Analysis:** ANCOVA is a statistical test that determines the significance of a set of independent variables on a dependent variable. This analysis would be used in order to determine the weights that should be assigned to the 4 factors in the complexity formula instead of using the same weight for all the factors.
- **Bayesian Modeling:** This model would better model the visual complexity of the Sankey Diagram. Unlike the current model, a Bayesian model is not a linear model. It computes a probability based on the features given that a diagram is a certain class

References

- Holtz, Yan, and Conor Healy. “Sankey Diagram.” *Sankey Diagram – from Data to Viz*, www.data-to-viz.com/graph/sankey.html.
- Chou, Jia-Kai et al. "Privacy Preserving Visualization: A Study on Event Sequence Data". *Computer Graphics Forum*. (2018).

FURI

Ira A. Fulton Schools of Engineering
Arizona State University