Visual Analytics of Multivariate Networks with Representation Learning and Composite Variable Construction

Supplementary Materials

A Supplementary Comparison of Scatterplot Designs (Sec. 2.3.3)

Scatterplot enhancements. Before we decided to design two-class density scatterplots, we tested various designs, including state-of-the-art enhancements. In Sec. 2.3.3, we showed differences between conventional density scatterplots and our two-class density scatterplots. Here, we compare conventional scatterplots, winglets, splatterplots, and two-class density scatterplots. In Fig. A.1, we show the results obtained by applying these designs to several preprocessed datasets used in our paper.

With winglets and splatterplots, it is still difficult to quantify the density at each location. This is critical for visually grasping trends and correlations. For example, from the splatterplot of Fig. 2.2-b2, analysts might misjudge `Class 0` has a higher density in general. Moreover, to effectively use winglets and splatterplots, there are various parameters we need to adjust. For example, winglets' appearance will be radically different based on the length of wings and the size of dots. On the other hand, splatterplots have several sensitive parameters, such as the threshold for contouring and the sampling rate for outliers.

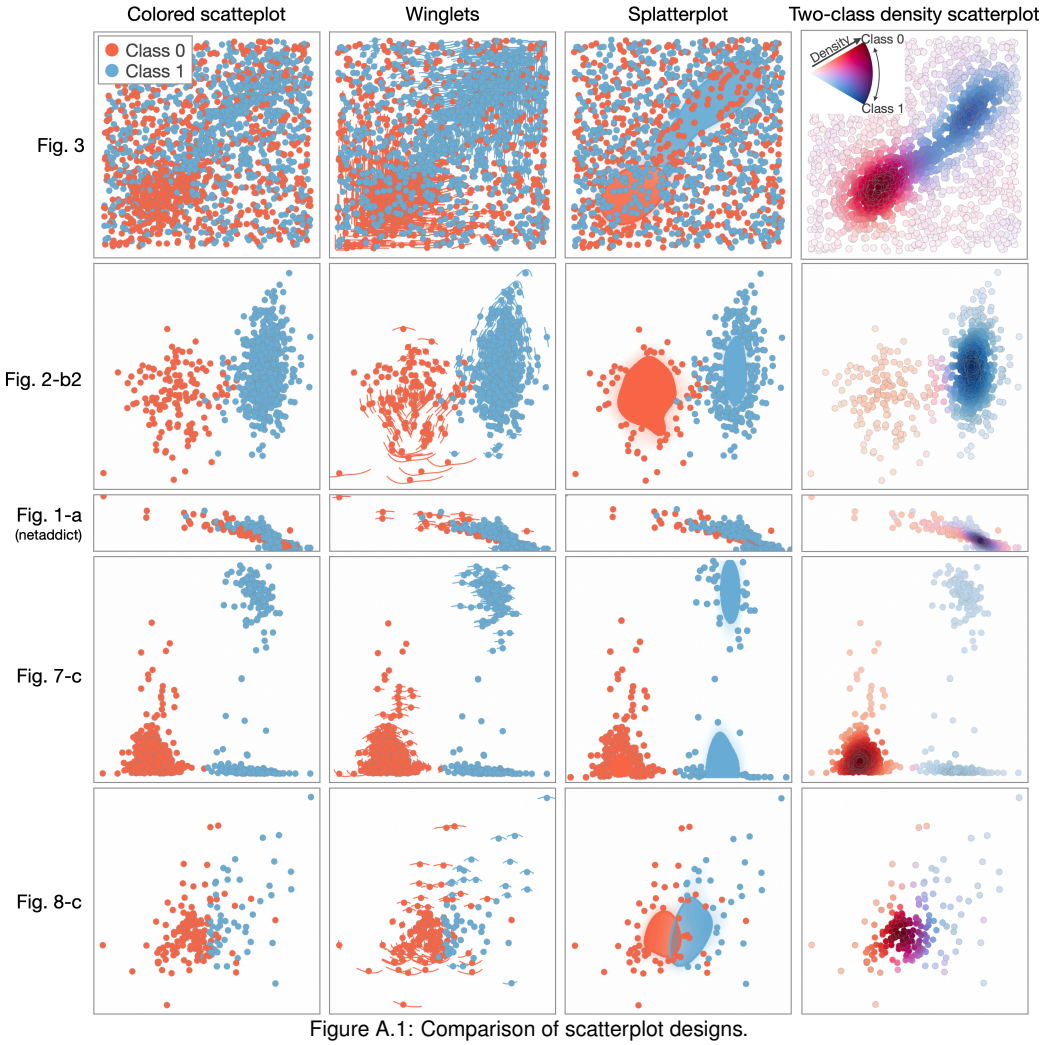

Figure A.1: Comparison of scatterplot designs.

Colormaps. Before deciding our final colormap design (Fig. A.2-d), we also tested various colormaps for two-class density scatterplots. For example, in Fig. A.2-a, we use Lespinats and Aupetit's bivariate colormap. This colormap uses hue to show the ratio of two attributes (i.e., the total density and the density ratio of two classes in our case) and lightness to represent the total value of the two attributes. From the visualization result, it is difficult to naturally grasp the two dense regions of `Class 0` and `Class 1` (located at the bottom left and the top right). As this issue is mainly caused by the inappropriate assignment of hue and lightness to the two attributes we use, the same problematic situation can be seen in Fig. A.2-b, where we use a different pair of hues (red and blue).

We can adjust Lespinats and Aupetit's bivariate colormap to assign hue and lightness to the total density and the density ratio, respectively. We can then show this colormap with a polar coordinate. As shown in Fig. A.2-c, this colormap can clearly convey each class's dense regions (dark purple and dark green regions). However, the colormap's hue is sensitive to which class has more density at each location. For example, in Fig. A.2-c, we observe a green region annotated by the arrow. This sensitivity is caused because the purple is located on the exact opposite side of the green in the CIELAB color space (note: even in other color spaces such as HSV). In our case, we are less interested in binarizing regions based on which class has a higher density. Rather, we want to show the quantity of the density ratio. Thus, we decided to use red and blue as two distinct hues (Fig. A.2-d), with which purple color is produced as each class's density becomes close to even. Based on this observation, we can create similar colormaps to Fig. A.2-d by avoiding selecting hues located on the opposite side in the color space. Fig. A.2-e shows such a colormap example, where orange and green are employed as two distinct hues (i.e., the mixed color becomes yellow). Both colormaps in Fig. A.2-d,e and similar ones are suitable for the two-class density scatterplot. Our implementation uses the one in Fig. A.2-d by default.

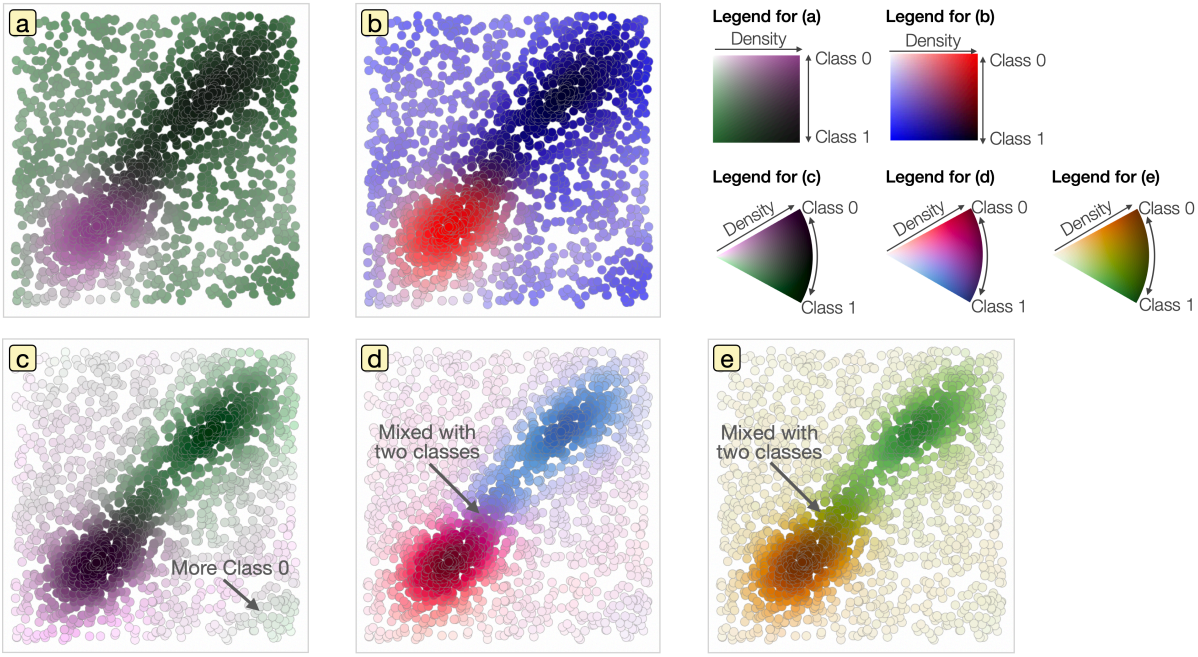

Figure A.2: Comparison of two-class density scatterplots with different colormaps.

B Additional Statistical Analyses on Composite Variables

Our workflow uses composite variables to help analysts interpret the learned network representations. Through the case studies, the effectiveness of the constructed composite variables is evaluated both qualitatively (e.g., visual patterns, expert feedback) and quantitatively (e.g., the improvement in correlation coefficients). Here, we further examine the effectiveness of the composite variables in the context of conventional analyses.

Case 1. We first visually examine data distributions corresponding to the original attributes and composite variable, as shown in Fig. B.1. We can observe that the composite variable, -0.4`open` + 0.7`consci` - 0.3`netaddict` + 0.3`min_happy` + 0.3`extra`, creates more clearly different data distributions between `Class 0` and `Class 1` than any single attribute.

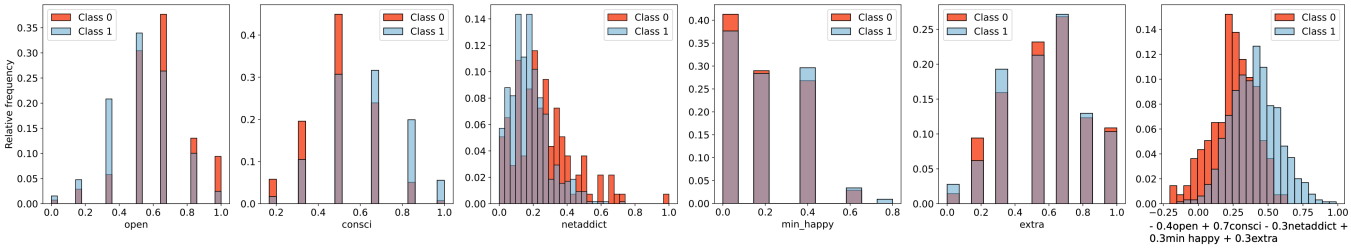

Figure B.1: Case 1: Relative frequency histograms for `Class 0` and `Class 1`. From left to right, sets of histograms correspond to `open`, `consci`, `netaddict`, `min_happy`, `extrac`, and -0.4`open` + 0.7`consci` - 0.3`netaddict` + 0.3`min_happy` + 0.3`extra`.

We next perform a set of computational statistical analyses: hypothesis testing of the difference between two groups, analyses of correlation between attribute values and classes, and classification using only one attribute/composite variable. For the hypothesis testing, we applied the Brunner-Munzel test because Levene's test implied we should not assume equal variances for several attributes (i.e., the Mann–Whitney U test is not appropriate). For correlation analysis, we computed the point-biserial correlation coefficient (r_{pb}) between the class IDs (i.e., 0 and 1) and attribute values. Lastly, for classification, we employed a linear discriminant analysis (LDA) with an oversampling method (specifically, the synthetic minority oversampling technique, or SMOTE) as `Class 1` has many more samples than `Class 0`. The results are shown in Table B.1. We can see the constructed composite variable provides notable improvements in correlation coefficients and classification accuracy.

	p-value (Brunner-Munzel test)	correlation coefficient (r_{pb})	accuracy (LDA)
`open`	*** 3.3×10^{-8}	-0.19	0.61
`consci`	*** 5.5×10^{-13}	0.24	0.64
`netaddict`	*** 2.7×10^{-7}	-0.24	0.62
`min_happy`	0.29	0.042	0.51
`extra`	0.99	-0.0020	0.49
-0.4`open` + 0.7`consci` - 0.3`netaddict` + 0.3`min_happy` + 0.3`extra`	*** 0.000	**0.32**	**0.67**

Table B.1: Case 1: Statistical analysis results. The bold fonts indicate considerable improvement by the composite variable.

Case 2. We repeated the same procedure as Case 1. The results are shown in Fig. B.2 and Table B.2. Fig. B.2 shows clear distribution shifts, where `Class 0` tends to have smaller values than `Class 1`. Also, all statistical measures showed large improvements achieved by the constructed composite variable. For example, only the composite variable shows the statistical significance difference between `Class 0` and `Class 1`.

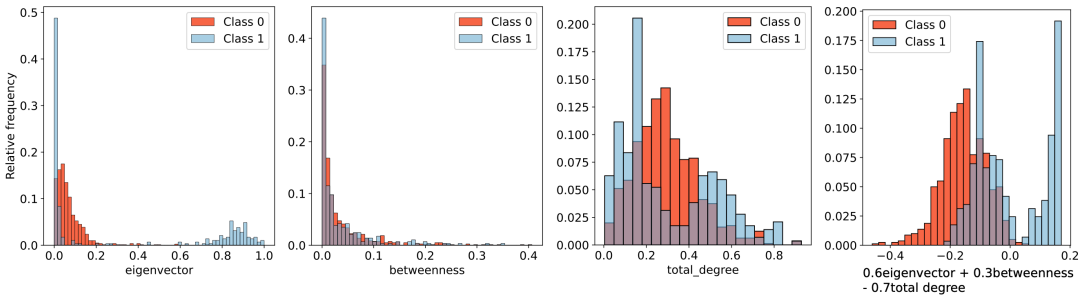

Figure B.2: Case 2: Relative frequency histograms for `Class 0` and `Class 1`. From left to right, sets of histograms correspond to `eigenvector`, `betweenness`, `total_degree`, and 0.6`eigenvector` + 0.3`betweenness` - 0.7`total_degree`.

	p-value (Brunner-Munzel test)	correlation coefficient (r_{pb})	accuracy (LDA)
`eigenvector`	0.074	0.47	0.67
`betweenness`	0.065	-0.020	0.51
`total_degree`	0.11	0.012	0.48
0.6`eigenvector` + 0.3`betweenness` - 0.7`total_degree`	*** **0.000**	**0.61**	**0.74**

Table B.2: Case 2: Statistical analysis results. The bold fonts indicate considerable improvement by the composite variable.

Case 3. We followed the same procedure as Cases 1 and 2. The results are shown in Fig. B.3 and Table B.3. Unlike the other two cases, we can only see subtle improvements (e.g., small improvements in correlation coefficients and classification accuracy). This situation of fewer improvements implies the difficulty of identifying the difference between the groups with simple statistical analyses. While the simplified representation in Case 3 utilizes neural networks and their nonlinear transformations (i.e., having high expressiveness), it still has a relatively low accuracy rate (0.82). A potential future improvement of the composite variable to aid in conventional statistical analysis is enabling the construction of more complicated composite variables, as described in Discussion section.

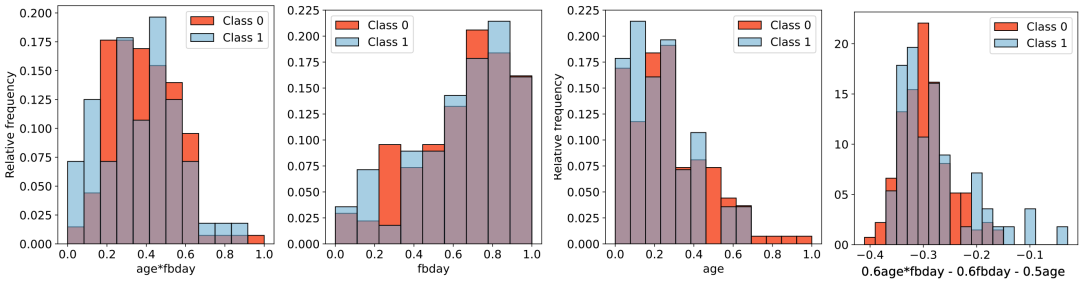

Figure B.3: Case 3: Relative frequency histograms for `Class 0` and `Class 1`. From left to right, sets of histograms correspond to `age*fbday`, `fbday`, `age`, and 0.6`age*fbday` - 0.6`fbday` - 0.5`age`.

	p-value (Brunner-Munzel test)	correlation coefficient (r_{pb})	accuracy (LDA)
age*fbday	0.74	-0.028	0.49
fbday	0.81	0.0090	0.51
age	0.11	-0.13	0.51
0.6age*fbday - 0.6fbday - 0.5age	0.50	**0.14**	**0.53**

Table B.3: Case 3: Statistical analysis results. The bold fonts indicate considerable improvement by the composite variable.

C Demonstration Video, Source Code, and Sample Data

We provide a demonstration video of interactive analysis with our workflow and UI. Also, to allow readers to test the workflow and UI, we provide related source code and sample data preprocessed from a publicly available dataset of faculty networks [89].

- Demonstration video: https://www.youtube.com/watch?v=Vro6uFGYBho
- Source code and sample data: https://github.com/hylu1994/Network-CV