Response to reviewer comments "NeuralNetTools: Visualization and Analysis Tools for Neural Networks" by M. W. Beck.

I thank the reviewer for providing thoughtful comments on my manuscript and the NeuralNetTools package. My response to these comments are shown in italics. Page and paragraph numbers refer to the original manuscript.

Reviewer C:

The manuscript identifies the need for open-source tools for visualizing the structure of a multilayer perceptron (MLP) neural network and for visualizing variable importance. The author provides an intuitive background of neural networks and a brief description of previous publications that develop tools for visualizing and understanding the MLP neural network. The supplied code works (mostly) as advertised (see below for discrepancies). Overall the paper is well written with a logical flow. The simple examples adequately demonstrate the tools. However, in my opinion the applied example does not fully draw-out the additional interpretation that is claimed by the NeuralNetTools package. Furthermore, there is limited usefulness in visualizing a neural network when there are more than about 10 input variables. (Although the same problem applies to visualizing a tree-based model when there are many terminal nodes.) Still I believe this tool could be useful for data sets with a small number of variables.

I certainly agree that plotting a neural network with many variables is impractical, but as noted, this is an issue with any plotting method for large structures. I believe the main benefit of the package is in the supporting functions, e.g., `garson`*,* `lekprofile`*,* `olden`*, that use the structure of the model (weights) and predictions to better understand relationships between variables. The* `plotnet` *function has importance as a basic visualization tool that informs additional analysis. Our hope is that this is clear from the existing text in the manuscript that immediately follows the plot section (i.e., first paragraph, sec. 3.2). I have added some text to the conclusions to re-emphasize this issue. First paraggraph of sec. 5: "Although visualizing a neural network with* `plotnet` *is impractical for large models, the remaining functions can simplify model complexity to identify important relationships between variables."*

The applied example section has also been revised to provide a more comprehensive demonstration of the utility of the package. Please see the specific response below.

I would recommend tentatively accepting the manuscript with the author addressing the following items:

• I would like to see a different applied example (or the current example changed) that gives a better practical demonstration of the ability to visualize a network and the corresponding interpretation.

Content was added to the section that shows the interpretation of a neural network model

using functions in the package. Figure 7 was added to show the results from a simple model with an interpretation provided in the text. Most of the addition to the text was included in the following paragraph (but see the applied example section for the entire revision):

"Figure 7 shows the information about arrival delays that can be obtained with the functions in ***NeuralNetTools****. The NID (7a) shows the model structure and can be used develop a general characterization of the relationships between variables. For example, most of the connection weights from input node I5 are negative (grey), suggesting that distance travelled has an opposing relationship with arrival delays. Figures 7b and 7c provide more quantitative descriptions using information from both the NID and model predictions. Figure 7b shows variable importance using the* `garson` *and* `olden` *algorithms. The* `garson` *function suggests time between destinations (*`air_time`*) has the strongest relationship with arrival delays, similar to a strong positive association shown with the* `olden` *method. However, the* `garson` *function shows arrival time as the second most important variable, whereas the* `olden` *function shows this variable as having a weak negative relationship with arrival time. This discrepancy is explained below. Finally, results from the* `lekprofile` *function (7c) confirm those in 7b, with the addition of non-linear responses that vary by different groupings of the data. Values for each variable in the different unevaluated groups (based on clustering) show that there were no obvious patterns between groups with the exception being group five that generally had early arrival times and long departure delays."*

• Abstract: please clarify what is meant by "unstructured datasets."

This was changed to "...among multiple variables".

• Introduction: I do not agree that "Data-intensive analysis is a relatively new research approach..."

The first sentence was revised: "A common objective of data-intensive analysis is the synthesis of unstructured information to identify patterns or trends 'born from the data'."

• Page 2, last paragraph: NeuralNetTools does not "improve the breadth and quality of information obtained from the MLP neural network." Instead, the package provides tools to better understand the MLP neural network.

The sentence was revised: "...that was developed to better understand information obtained from the MLP neural network."

• Page 6: I am not familiar with the concepts of power and degrees of freedom with neural networks. Please clarify the sentence, "Pruning has additional benefits..."

These sentences were clarified to emphasize that pruning can aid in 1) the interpretation of relationships between variables by removing unimportant connections, and 2) successive model fitting by reducing the number of weights that are changed during model training: "In addition to visualizing connections in the network that are not important, connections that

are pruned can be removed in successive model fitting. This reduces the number of free parameters (weights) that are estimated by the model optimization routine, thereby increasing the likelihood of convergence to an estimable numeric solution for the remaining connection weights that minimizes prediction error (i.e., model identifiability, Ellenius and Groth 2000)."

• Page 8, Section 3.3: It is not clear to me how the Lek profile is used to evaluate interactions between variables.

The `lekprofile` *function evaluates the change in the response variable across the range of values for a given explanatory variable while holding the remaining explanatory variables constant. I have considered these plots conceptually similar to traditional interaction plots (i.e., what is the effect of an explanatory variable at different values of another explanatory variable), but I suppose there is not a 1:1 correspondence. The* `lekprofile` *function is more appropriately used to evaluate the shape of the response for a given explanatory variable, as opposed to the importance of variables as for the other functions. I've replaced 'interactions' with 'relationships' as this description is more appropriate.*

• Figure 6: I was not able to reproduce this plot. In addition, when I generated this plot, the bars for two of the groups contain zero (see Groups 1 and 2) - that is, the bar is greater than zero for part of the bar and is less than zero for the other part of the bar. I thought these bars would be bounded above or below by zeroperhaps I'm not understanding this plot.

This was an issue with stacking of non-positive values in ggplot bar plots. The situation has been resolved. Additionally, the default position of the bar plots is now `'dodge'` *(side-by-side for each group), as passed to* `ggplot2::geom_bar`*. This was implemented in the development branch on the GitHub page and will be pushed in the next release to CRAN.*

• Section 4: When I ran the code provided in the git repository, I could not reproduce Figure 7 (see attached Figure7.png file). I added the following line before the toplo1 line to make this work:

```
imp_sums$variable
```

Thanks, this was fixed.