Paul de Fusco - DSE241 - February 15, 2018

Exercise 5 Report: Sheep Experience Exploration and Visualization

Objective:
Although a small dataset, the Haas-Bighorn Sheep Dominance dataset presents with opportunities for network exploration and feature engineering aimed at uncovering insights into sheep behavior and dominance. The creator made four visualization idioms to explore the datasets and the correlation between sheep dominance and variety of sheep interacted with.

Task:
The main question the author answers with his visualization is: is sheep dominance correlated with the amount of different (unique) sheep it interacts with (both when it was dominated or it dominated another sheep)? The creator intends each idiom as a “Step”. Each addresses a subtask providing further insight leading up to the final answer to the task defined above. The author conceived the four step visualizations to be visualized in counterclockwise order and starting from the top left quadrant:

1) Step 1 Visualization: the author addresses the subtask of providing an introductory visualization by showing the more dominant sheep and how they relate to other sheep in a spectral layout. Specifically, the author shows that some sheep have been more dominant than others, which sheep have been dominated the most, and to which sheep each sheep dominance directs towards to. However, while many of the dominant sheep (largest circles and deeper blue) are both dominant in terms of percentage of total dominant interactions (circle size) and the propensity to dominate (blue intensity), a few showcase only one or the either. This indicates that they are not truly the most dominant in the network: for example they may simply have had fewer interactions but overall often dominated throughout them. The measures of dominance are all shown in the legend in the graph and are:
   A. Circle Size: larger circles reflect more dominance defined as “Count of Dominance Weight for Particular Sheep by Percentage of All Sheep Interactions”
   B. Circle Blue Intensity: higher intensity reflects “Ratio of Count of Times a sheep dominated vs. was dominated”
   C. Circle Border Width (in red): higher border width reflects higher “Count of Interactions in which Sheep was Dominated”
   D. Edge Thickness (interactive): by default drawn as gray dotted lines between nodes with higher thickness reflecting higher weight for the edge. When hovered on, the edge color reflect weight on a spectrum of light green to dark red for lowest to highest; in addition, lines turn from dotted to solid.

2) Step 2 Visualization: the author addresses the subtask of providing the same information as in visualization 1 but in a more user-friendly way by using a shell layout to arrange sheep and graph edges. By looking at this idiom, the viewer is more quickly able to spot weak and dominant sheep. Additionally, the layout allows to more easily visualize edges for a particular sheep and notice some sheep dominate particular targets more frequently and vice versa. Although the information is color encoded in the same way as in Step 1 and no additional dataset features are shown, this visualization is key to introduce Step 3 and allow easy comparison with it.

3) Step 3 Visualization: the author addresses the subtask of introducing attributes on variety of sheep interactions, while encoding these in a layout that is similar to the one in Step 2 in order to provide an easy way to compare how the new features relate to dominance.

Specifically, the author shows that sheep that are more dominant in terms of percentage of dominant interactions are generally also the ones that have had more interactions with a wider variety (unique) of sheep as opposed to the ones who have only interacted with the same few sheep. However, he also demonstrates that while some older sheep have a higher dominance percentage index, there are many exceptions. Furthermore, the viewer can see that sheep age does not necessarily correlate with its propensity to have had higher interactions with different sheep. The metrics are shown in the legend and are:

A. Square Size: larger squares reflect higher sheep age and vice versa
B. Inner Square Color (Green, Interactive): higher green intensity reflects more interactions with different sheep; to view, hover over square
C. Square Border Color (Orange): higher intensity of orange implies higher propensity to be dominant during interactions calculated as percentage of dominant interactions to total interactions for the specific sheep (not the whole network as in the case of the blue intensity in visualization 1 and 2)

Note: Step 2 and 3 visualizations are coupled interactively and laid out in identical fashion in order to allow easy comparison and enable the viewer to easily spot dataset peculiarities. Thus, by looking at both visualization steps concurrently, new joint subtasks are addressed:

I. There is direct correlation between the propensity of sheep to dominate (blue intensity in left visualization) as many of the weakest sheep are also the youngest, but there are also exceptions e.g. sheep 3,4, 5 and 22.
II. The youngest sheep are dominated by particular sheep more frequently than others (while most edges are green and yellow, often only one connecting to these sheep is in red). In the contrary, more dominant sheep have a wider variety of red edges connecting them to other sheep.
III. However, despite what stated in #2 above, many of the dominant sheep do not have a much higher count of edges connecting them to different sheep, indicating that most of their dominance is directed towards the same few targets. ***This information is a key takeaway leading us to Step 4***

4) Step 4 Visualization: the author addresses the subtask of rearranging the information in a scatter plot to highlight the insights gained from comparing step 2 and 3 idioms. On the x-axis the author displays the age, while on the y axis he displays the count of total interactions with different (unique sheep) - both dominated and dominating. The color of the circles reflects the square border color from step 3 i.e. the propensity to be dominant during interactions calculated as percentage of dominant interactions to total interactions for the specific sheep.
   A. The task demonstrates that while dominant sheep tend to scatter to the right of the plot reflecting a higher age (as shown earlier this correlation has exceptions) the number of different sheep a particular sheep has interacted with does not seem to play a factor at all as there isn't much of an inverse or direct relationship between the x and the y axis. In other words, while color depth increases with age (x axis), it does not vary with respect to interaction variety (y axis).

Data Augmentation:
The author created new features to derive more insights and added them to the nodes dataset:

1) DominatedBy: count of how many times any sheep dominated the particular sheep
2) DominatedOthers: count of how many times any sheep was dominated by the particular sheep
3) DomBy/DomOthers: ratio of 1 and 2
4) TargetsPerSource: count of how many unique different sheep the particular sheep dominated
5) SourcesPerTarget: count of how many unique different sheep the particular sheep was dominated by
6) DominatedOthersPCGT: percentage of DominatedOthers across DominateOthers sum across all sheep
7) TotalInteractions: sum of 1 and 2
8) WonPCGT: percentage of interactions in which sheep dominated (calculated with only that particular sheep total interactions as denominator, not the network total)

9) tot_unique_interactions: sum of 4 and 5 (for each sheep)

Discussion on Expressiveness, Effectiveness, and Color:
The four visualizations as a whole respect the expressiveness principle as they encode all the information contained by the dataset, including augmentation features. No additional information is shown and all inherent data magnitude is conveyed on an objective scale by its salience in the graph context. Specifically, the author uses a combination of color features and object size to convey different levels salience for each node and edge. For example, nodes that are more dominant are bigger and bluer, making them very easy to recognize.
In more detail, channel effectiveness criteria are explained below:

1) Accuracy: dominant nodes as well as more frequent interactions between the same sheep are depicted by increasing circle/square size and edge thickness according to a linear scale of magnitude: the bigger/thicker the object, the greater the value. While a nonlinear scale could increase standout for particular nodes showcasing more dominance, the idioms would suffer from circles and rectangles that are too large. The magnitude of interactions is also reflected by edge color intensity, with higher values in a dark red (a more noticeable and intense color) and lower values in lighter and lighter green. Yellow reflects intermediate values in the spectrum.

2) Discriminability: in step 1, most of nodes are at first compacted in the same area. This is a wanted effect to render a true mathematical visualization of the network (which also has an aesthetically pleasing aspect to it). The viewer can however easily distinguish the nodes by zooming in and moving the cursor, thus making every node distinguishable. With respect to idioms 2 and 3, the viewer can easily discern between different nodes and idiom channels as a result of the circular layout. In step 4 the viewer is presented with only 28 scatter points across a large space, with very little overlap. The zooming and pan tools are also available in Steps 2, 3, and 4.

3) Separability: the creator was able to add a few visual channels to each idiom without overcrowding each of them and impair visibility. Nodes are easily separable with the provided tools (when needed). Each node encodes 3 channels in each idiom. While in Step 4 the use of a scatter plot makes risk of separability issues low, these problems are avoided by the creator in Steps 1-3 by using scales of magnitude that are consistent with visibility with respect to each channel. In addition, hover tools make showcase additional information only when specific objects are highlighted, e.g. making it easy to separate a particular edge from the rest when the cursor is moved on it.

4) Popout: colors are combined to produce more visually effective results. For example, idiom 1 and 2 use a combination of blue and red, so to allow the viewer to clearly notice the border (red) which is not as large of a presence on the idiom in terms of size. Also, the author uses a dark red to convey a sense of weakness and “pain” in relation to the circle border channel. This provides the viewer with the intuition that the more noticeable the red, the weaker the sheep evaluated.

5) Grouping: nodes are by definition grouped in the same network and related to one another by edges. However, idiom 1 gives a more spatial representation of grouped nodes. Nodes that are closer to the center are more dominant and have more interactions with more nodes. Nodes on the periphery are less interactive. Steps 2 and 3 don’t use position to group nodes, but step 4 uses a scatter plot to group nodes that showcase the same magnitude in terms of the x-y axis features.