Exploration of Multi-Network Data with Heatmaps

Concordia University

Department of Mathematics and Statistics

ISM Undergraduate Research Report

Phil Boileau

Supervisors:

Dr. Lisa Kakinami

Dr. Lea Popovic

September 2017

Acknowledgments

I'd like to thank Dr. Lisa Kakinami and Dr. Lea Popovich for their encouragement and guidance throughout the summer, as well as for giving me the opportunity to complete this project. I'd also like to thank the Institut des sciences mathématiques for their support, and Dr. Tracie Barnett for granting us access to QUALITY's social network data.

Introduction

Whenever a researcher acquires new data, their initial reaction should be to perform an exploratory data analysis because "[it] isolates patterns and features of the data and reveals these forcefully to the analyst".[1] Traditionally, this critical step has been accomplished with simple visual aids, such as scatterplots, boxplots and residual plots. However, these tools become inefficient and are unable to describe all aspects of the data as datasets grow increasingly complex. There is mounting evidence indicating that for the creation of successful visualization methods, a combination of statistical and interactive techniques is essential.[2]

One field of study that has been benefiting from recent innovation in exploratory data analysis methods is network research. Networks allow scientists to investigate the structure of the systems they are researching, which in many cases offer more insight than the analysis of only their individual components.[3] For example, epidemiologists use networks to study the spread of disease in various populations across the globe. By simulating the underlying structure of the social networks within these populations, they can identify behaviours that contribute to the propagation of illnesses.[3,4]

Simply put, a network is a set of points that may or may not be linked together by vertices[3]. In general, the points are referred to as nodes and the links are called edges. They have been used to model all kinds of systems in a plethora of disciplines, from the structure of the world wide web in information science to the representation of the chemical interactions that fuel cells and organisms in biology.[3]

The examples above demonstrate the wide variety of systems that networks can be used to model. They range from the simplest of structures (such as a single node), to the unbelievably large, such as the world wide web with its billions of nodes and edges. However, network topology can be described by much more than just these points and links. For example, the level of connectedness between the elements of a network is described by it's density, the importance of specific nodes may be categorized by it's centrality and the degree to which similar nodes are grouped together is designated by the network's homophily. These structural characteristics offer researchers insights on the inner workings of the systems they study.

Conceivably, exploring a network dataset can be a daunting task; not only does the data typically have large dimensions, it is a mix of four types of variables: graph, node, edge and topological. Graph variables describe the general characteristics of the network, node and edge variables describe the attributes of the nodes and edges and topological variables detail the structure of the network. Take for example a friendship network in an elementary school class where nodes depict children and edges represent friendship ties between classmates. In this model, the grade level of the class is a graph attribute, the gender of each child is a node characteristic, the strength of friendship between children is an edge characteristic and the number of kids in the class is a topological variable.

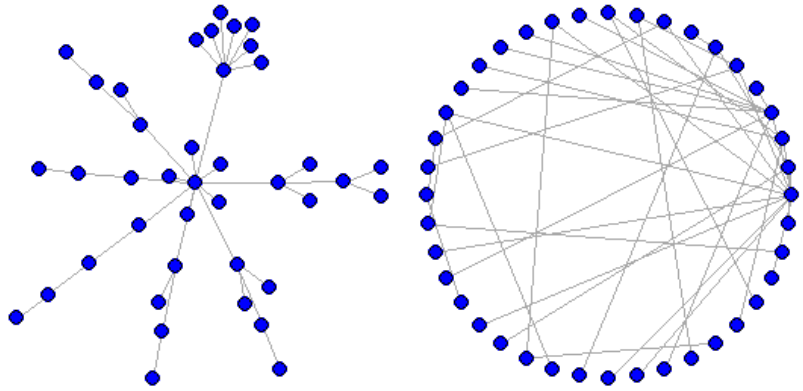

Figure 1: Two different node-link diagrams of the same network. There are a multitude of methods to visualize a network, and each one has its strengths and weaknesses. The graph to the left facilitates the determination of its diameter (i.e. the longest path between two nodes), and the one on the right illustrates its connectedness.

Luckily, there are software packages such as Gehphi, Igraph and NodeXL that have been developed to analyze and create visual representations of networks. The most common plotting technique for such data is the node-link diagram **(fig. 1)**. Gehlenborg and Wong state that these illustrations make it easy to discern some topological qualities of the network, such as finding the nearest neighbours for a specific node or tracing paths between its elements.[5] Note that the node-link diagram is not restricted to the display of structural qualities**.** Unfortunately, this method is limited by the size of the network: as the number of nodes and edges increases, the visualization becomes cluttered and illegible. Furthermore, there amount of information that can be presented in these diagrams is constrained. It is impossible to get an overview of all the variables comprising the network with the node-link illustration.

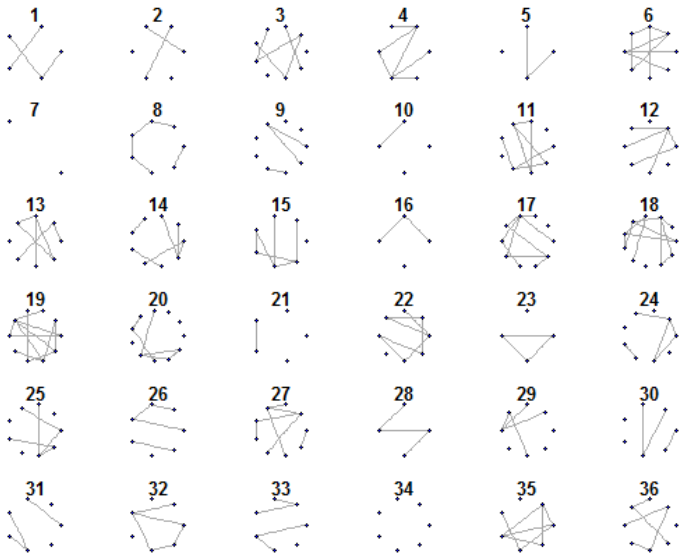

Figure 2: The node-link diagram of 36 networks from a simulated multi-network dataset.

The issues faced when using this visualization method are compounded when attempting to perform exploratory data analysis on multi-network data, which is composed of multiple disjoint networks that share common graph, node, edge and topological variables. Not only is it cumbersome to display a sufficiently informative portion of the graph, but simple comparisons between graphs, such as the number of nodes or edges, are difficult **(fig. 2)**. This visualization method is ineffective at depicting relationships between variables and networks and at describing the distribution of the characteristics.

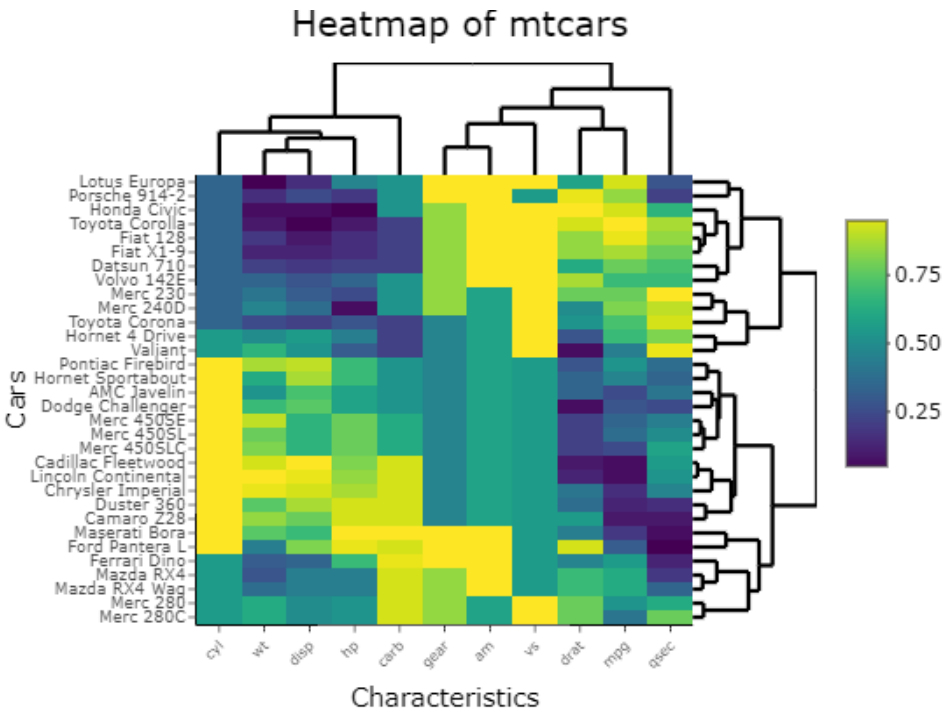

Figure 3: A heatmap of select cars from 1974 and their various characteristics. Data provided by "mtcars", a popular R data set.

An alternative representation technique that may be more appropriate for multi-network data than the node-link diagram is the heatmap, a popular plotting method in biology used to represent multivariate data (such as gene expression matrices).[6] A heatmap **(fig. 3)** is a colourful graphical representation of a two-dimensional matrix.[6] Heatmaps rely on two key elements to convey legible information to their users. (1) Before the multi-network data is ready to be displayed in the form of a heatmap, both the networks and the collections of their attributes (graph, node, edge and topological) must undergo hierarchical cluster analyses. This step insures that the colour matrix is well organized. (2) Additionally, the appropriate color palette must be chosen to represent the data in the heatmap.[6] If the goal is to highlight a single extreme, then the use of a single-color gradient is suitable. On the other hand, if the objective is to emphasise both extremes, a divergent color gradient of three hues are required.[6] These features maximize the pattern recognition potential of the resulting graph, which is the main goal of any data visualization. They can display large quantities of information in a single figure, making them the ideal tool for presenting multiple networks along with their graph, node, edge and topological variables of interests.

The objective of this research was to demonstrate how heatmaps can be a useful tool for the representation of multiple network data. In particular, the utility of heatmaps in (1) demonstrating relationships among the graph, node and structural characteristics of the networks simultaneously, and (2) illustrating the distribution of variables across networks in order to enable comparison of various elements of the data were of interest. All analyses were conducted with R.

Methodology

Data Set

Data for this study were from QUALITY cohort. The QUALITY study is a longitudinal investigation into the causes and consequences of obesity in children and teenagers in Montreal (n=630).[7] For the purposes of this study, the pilot social network dataset collected in a subsample of the cohort (n=46) was used.

Neatmaps Package

To streamline the process of exploring multi-network data with heatmaps and to insure reproducibility, the R package *neatmaps* was created. Documentation is available the CRAN website and on GitHub. Starting with the raw multi-network data, the program calculates the topological characteristics of the networks, prepares the data for analyses, performs hierarchical clustering on the networks and their attributes (including bootstrap validation, see **Statistical Analysis**) and then generates multiple heatmaps. Additionally, a template with instructions is furnished to users who wish to create dynamic reports of the results of the exploratory analysis. All analyses performed on the QUALITY social network data relied on *neatmaps*. This package depends heavily on the *Igraph*, *pvclust*, *heatmaply* packages.

Heatmap Feature: Hierarchical Clustering

Hierarchical clustering is a group of unsupervised learning techniques that aims to find clusters in a data set. This method divides the data into subgroups based on the similarity of elements in the groups. The most popular hierarchical clustering method (and the one used in the analysis of the QUALITY data set) is the bottom-up technique.[8] The name is self-descriptive. Each element of the dataset initially forms it own cluster (bottom), and is paired with other clusters based on their similarities, hence forming a new group. As the process continues, clusters increase in size until all the subgroups form a single mass (top). The resulting visualization is an upside-down tree that is called a *dendrogram* **(fig. 4)**. The algorithm to perform this technique is as follows[8]:

1. Begin with a set of n elements. Let each observation be an individual cluster. For all $\binom{n}{2}$ pairs of vectors, compute the pairwise similarity.
2. For $i = n, n-1, \ldots, 2$
    a. Inspect all pairwise similarity measures among the clusters and join the clusters that are most similar. The level of similarity between two groups dictates the height of their fusion in the dendrogram.
    b. For the remaining $i-1$ clusters, calculate the $\binom{i-1}{2}$ measures of pairwise similarity.

One of the most common R functions used to perform hierarchical clustering is *hclust()*, which belongs to the *stats* packages.

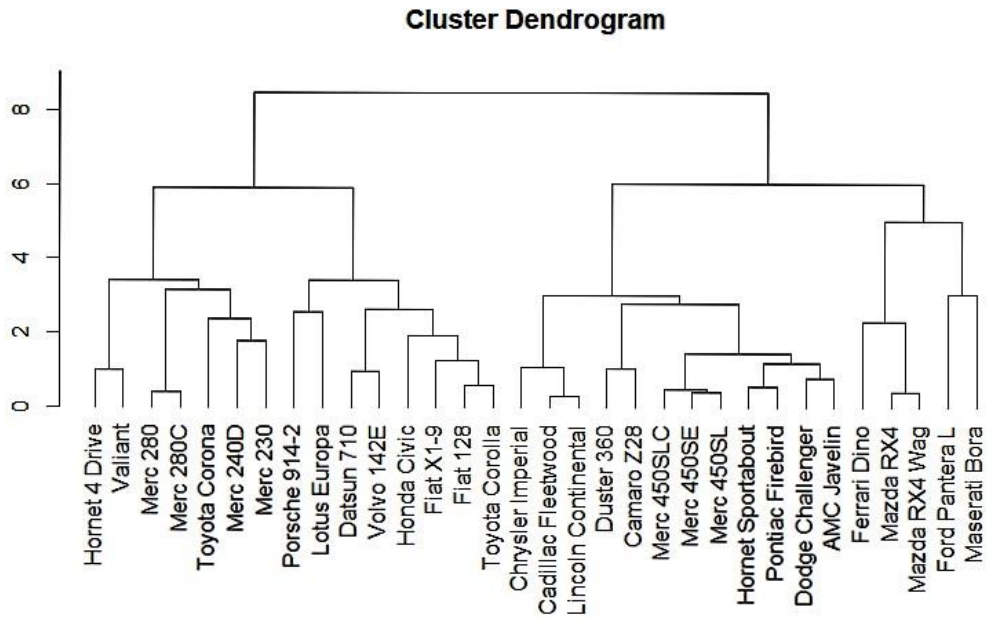

Figure 4: A dendrogram of depicting the hierarchical clustering of the mtcars data set.

As mentioned above, elements of the data set are clustered based on their similarities. There are many ways to measure this, but the most common approach is to minimize the Euclidean distance between the clusters. Since each item of the data set can be represented by vectors of identical size, the distance between two elements of length m can be calculated as follows:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_m - b_m)^2}$$

This method for determining the level of similarity between clusters may seem, but there are important aspects of the data to consider before using it to perform hierarchical cluster analyses. These details include, but are not limited to, (1) the appropriate preparation of the data to ensure that distances between elements of the data set are measured accurately and (2) selecting the appropriate technique for computing the distance between clusters as they grow to contain multiple vectors.

Data Preparation

Because of the use of Euclidean distance to measure the level of similarity among groups during hierarchical cluster analyses, the data being studied must be: (1) scaled, (2) quantitative, and (3) using an appropriate linkage method.

Scaling

To ensure that all potential relationships carry the same weight, data must be scaled. Take for example the numerical vectors A, B and C that represent a set of characteristics belonging to a group of people. The range of the elements of A is $[0,1]$, the range of the elements of B is $[0,0.1]$ and the range of the elements of C is $[0,100]$. Then $d(A,B) < d(A,C)$, regardless of the underlying relationship between the characteristics.

There are many ways to scale the data. One of the most common methods is to normalize the data, only to be used when it is assumed to be normally distributed. The computation is as follows:

1. Let D be a matrix containing the data with n rows and m columns.
2. Let Z a matrix with the same dimensions as D.
3. For each column j in D:
    a. For each row i in D:

$$Z_{i,j} = \frac{D_{i,j} - \overline{D_j}}{S_{D_j}}$$

The base library of R contains a function, *scale()*, for performing this operation on numerical data sets.

Another common technique for scaling data sets is to use the empirical distribution function,[9] which transforms the elements of a vector into their percentile values. This scaled matrix is computed as follows:

1. Let D be a matrix containing the data with n rows and m columns.
2. Let P a matrix with the same dimensions as D.
3. For each column j in D:
    a. For each row i in D:

$$P_{i,j} = F(D_{i,j}) = \frac{number\ of\ elements\ in\ column\ j < D_{i,j}}{n}$$

This approach transforms the ranges of each column of the data set to $[0, a_j]$, where $0 < a_j \leq 1$. One of its weaknesses is its inability to transform the ranges of the data matrix such that they are all equal. Consider a column which happens to have a quarter of its elements equal to its largest value, the range of the transformed column would only be $[0, 0.75]$. This reduces the accuracy of the similarity computation, the Euclidean distance between clusters, and therefore affects the quality of the hierarchical clustering.

A third procedure that improves upon the shortcomings of the previous two methods rescales the data set such that each column has a range of $[0, 1]$:

1. Let D be a matrix containing the data with n rows and m columns.
2. Let S a matrix with the same dimensions as D.
3. For each column j in D:
    a. For each row i in D:

$$S_{i,j} = \frac{D_{i,j} - \min(D_j)}{\max(D_j) - \min(D_j)}$$

This method preserves the underlying distribution of each column of the data and forces the range of each column to be identical. For these reasons, this scaling technique was chosen to analyze the QUALITY social network data. For the remainder of the report, it will be referred to as the *rescaling* method.

Quantitative

Another important aspect to consider before performing a hierarchical analysis are the types of variables that compose the data. When using Euclidean distance to measure similarity, categorical data should be avoided as distances between these kinds of variables are meaningless. The same can be said for binary data, or any quantitative data displaying negligible variability.

Linkage Methods

An important facet of hierarchical clustering is selecting the appropriate cluster comparison technique. There are many linkage methods, and their goal is to maximize the similarity between sets.

Linkage Method	**Description**
Complete	Minimizes the largest distance between the elements two clusters. Dendrograms tend to be balanced and less susceptible to noisy data and outliers. However, large clusters may be split unnecessarily as it is biased towards same sized clusters.
Single	Minimizes the minimum distance between the elements of two clusters. Using this method can result in extended, chaining clusters and it is sensitive to noise and outliers.
Average	Minimizes the mean distance between elements of two clusters. It is a mixture of the complete and single linkage methods.

Table 1[8]: A summary of the most commonly-used linkage methods.

The linkage method chosen can have great effects on the results of the hierarchical clustering. James et al. therefore recommended performing analyses with all three techniques to discover which patterns consistently emerge from the data.[8]

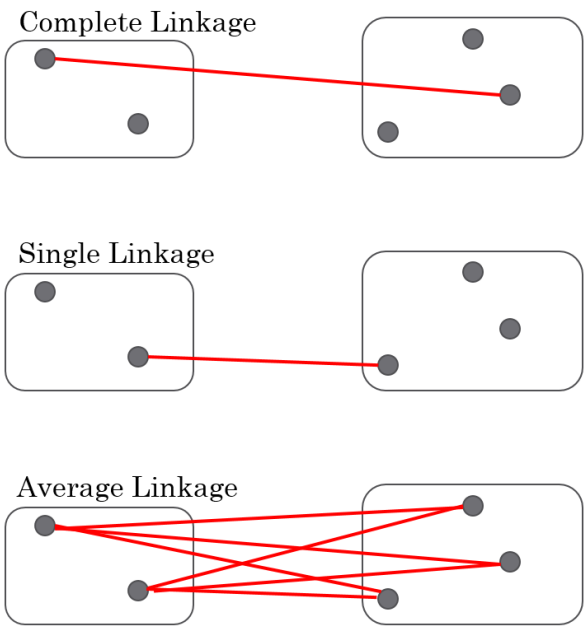

Figure 5: A visual representation of the linkage methods presented above. Each box represents a separate cluster, and the line(s) joining the clusters together illustrates the element of the groups involved in their fusion.

Statistical Analysis

Validating the results of cluster analyses is an important step in determining whether the subgroupings in the data are meaningful or if they simply constitute noise. There are many different techniques available, but there is no consensus on which is best. To test the legitimacy of the clusters found in the QUALITY social network data, the bootstrapping method was used.

The bootstrapping method consists of performing the hierarchical cluster analysis for a predetermined number of iterations, where each iteration has a few elements of the data set removed and others are duplicated such that the number of elements being clustered is fixed. If a certain group of elements tend to cluster together for many of the iterations, then the cluster is significant and likely indicates a relationship among its elements. This metric is called the bootstrap probability (BP). Take for example the first subgroup of cars in figure 4, the Hornet 4 Drive, the Valiant, the Mercedes 280, the Mercedes 280C, the Toyota Corona, the Mercedes 240D and the Mercedes 230. Imagine the bootstrap method was applied to the mtcars data set for 1000 iterations, and that the cluster was given a BP of 65%. This signifies that the elements of the cluster grouped together for 650 of the 1000 iterations. It is worth noting that the topology

within the cluster probably varied over the course of those 650 trials. Unfortunately, the BP statistic is biased, meaning that a low value does not necessarily indicate that a cluster is insignificant.[10]

Shimodaira developed a statistic called the approximately unbiased p-value (AU) that can be interpreted in a similar way to the BP.[11] This is the result of a multistep-multilevel bootstrap algorithm, which means that instead of keeping the sample size of elements fixed in each iteration, as is done with the computation of the BP, the sample size varies between iterations. An AU value larger or equal to $1 - \alpha$ should lead to the rejection of the null hypothesis, which assumes that the cluster is insignificant.

The R package *pvclust*,[12] created by Shimodaira and Suzuki, simplifies the process of determining the significance of groupings found during a hierarchical cluster analysis. This is accomplished either by visual means that rely on dendrograms which highlight significant clusters **(fig. 6)** or by way of tables. This package was used to measure the significance of the clusters found in the QUALITY social network data. The dendrograms produced by the clustering are then used to order the elements on axes of the heatmap.

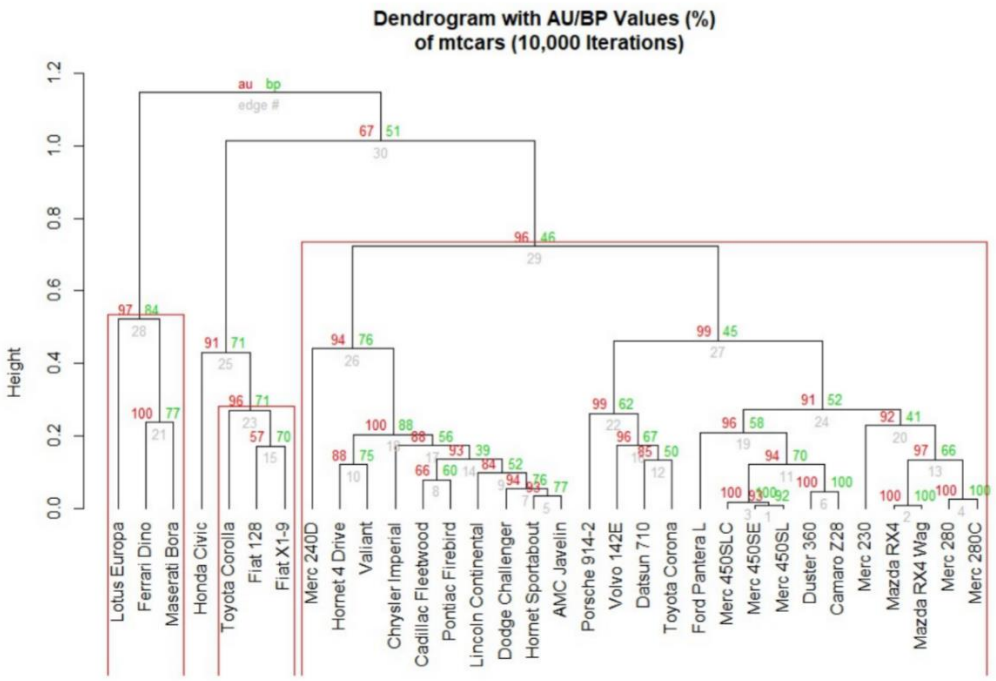

Figure 6: A dendrogram produced with the *pvclust* package that depicts the significant clusters of the various cars in the mtcars data set. Notice the differences between the AU and BP values, in red and green respectively.

Case Study - QUALITY Social Network Data

The Quebec Adipose and Lifestyle Investigation in Youth (QUALITY) is a longitudinal study designed to investigate the natural history of cardiovascular disease and type 2 diabetes in Quebec youth.[13] The original study consisted of over 600 children who are at an increased risk for obesity (due to having at least one obese parent, as per the eligibility criteria). Data collection is ongoing in both children and their parents.

For this study, data from a pilot project revolving around the collection of social network data for 46 of the QUALITY study participants were utilized. Each of the chosen youths, called egos, was asked to list up to ten of the most important people in their lives, called alters. Study participants then answered questions pertaining to the demographic characteristics, physical activity habits, the eating habits and the body images of these people. The participant also had to quantify, in terms of closeness and importance, their relationships with these individuals, as well as describe the relationships among the listed persons. With this information, 46 social networks were constructed. A combination of heatmaps and dendrograms were used to illustrate the significant relationships among the various variables of the social networks.

Results

Not all the networks and the variables of the QUALITY social network data were included in the analysis, either because (1) they were not appropriate for hierarchical clustering, (2) the information they provided was redundant or (3) they had missing entries. Thus 35 social networks and 43 of their shared characteristics were analyzed. Since the data were not normally distributed, the data were transformed using the rescaling approach introduced in the data preparation component of the methods section. Euclidean distance was used to measure the similarity the node attributes and networks clusters.

Heatmaps using the single, average and complete linkage methods for the network axis demonstrated no major differences between linkage methods. For this reason, the network axis for all the following heatmaps where hierarchically clustered using the complete linkage method. In contrast, the variations of the heatmaps (and their respective dendrograms) using the single, average and complete linkage methods for the network attributes are presented. For comparison purposes, a node link diagram of the same data is also presented. When comparing the heatmaps and the dendrograms produced by the linkage methods, three things are immediately apparent.

First, all three heatmaps have columns of primarily dark blue and red cells, which indicates that there are network variables in the data that exhibit extremely little to no variation in their values. This signifies that although these variables are clustering with others of the same colour, it may not be caused by an underlying relationship between them, but because the variation in their values is extremely limited, affecting the similarity calculations. These variables are: the weekly duration of time the ego spent on the internet for entertainment (blue) and the frequency the ego would start a diet for weight loss purposes (red). To remedy this, the questionnaire used for data collection should be modified to ensure a wider range of responses for these questions.

Second, the dendrograms of the network variables produced by the three linkage methods are quite different. The single linkage method formed an elongated chain **(fig. 9)**, indicating that outliers in the variables are affecting the hierarchical clustering. The complete linkage method forced the variables into clusters of similar size **(fig. 13)**, whereas the average linkage method formed groups of varying size **(fig. 11)**. The differences between the hierarchical clustering

outcomes suggest that the average linkage methods dendrogram paints the best portrait of the network variables' relationships.

Third, the significant clusters determined by each of the linkage methods vary in size and/or contents. Of the two significant clusters found using the single linkage method **(fig. 9)**, only the first can be construed. The second, consisting of 30 variables, is much too large to be interpreted; the cluster cannot be labelled as a family of variables (e.g. physical activity variables, social variables, etc.). On the other hand, the significant clusters found by the complete linkage method **(fig. 13)** are much smaller than what common sense and the literature would have expected. For example, the clustering of the ego's fat-mass percentage and BMI are not significant. Once again, the results of the average linkage method **(fig. 11)** are the most reliable: the clusters are small enough to be interpretable, but large enough to offer insight on the variables' relationships. This leads us to conclude that the average linkage method has produced the most dependable heatmap **(fig. 10)** for the exploration of the QUALITY social network data. Having selected the most appropriate heatmap, the interpretation of the visualization can now take place.

The first significant cluster consists of the number of people in the participant's network and the alters mean number of ties within the network **(fig. 11)**. In the heatmap, the adjacent cells in each variables' columns tend to have an identical or similar colouring, meaning they have similar values. This can likely be explained by the fact that participants nominated alters originating from the same social circles. For example, if an ego's social network consists of 5 family members, the expected number of ties for each relative would be 5 as well.

The second significant cluster can be interpreted as a grouping of lifestyle variables that affect the ego's fat-mass. In particular, the frequency of physical activity of the alters, the gender of the alters and the eating habits of the alters have all been discovered to impact the ego's physical activity, TV watching and eating habits. The relationship between these behaviours and childhood obesity is documented. [14–16] The heatmap depicts these complex relationships and their underlying distributions, which can lead to the formulation of new hypotheses or to the validation of established results.

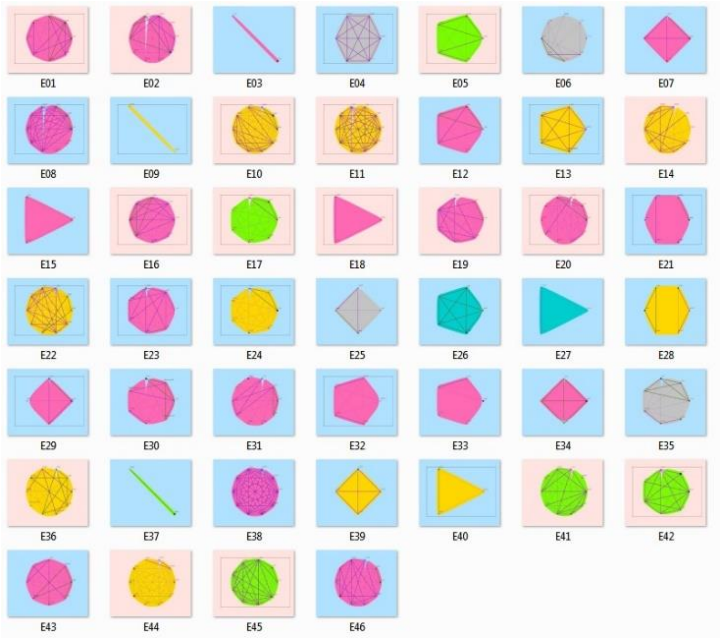

Figure 7[17]: The node-link diagram of 46 ego networks from a multi-network dataset. The background colours represent gender of the ego and the colours used to highlight the network indicate the range of the ego's BMI.

Single Linkage Method:

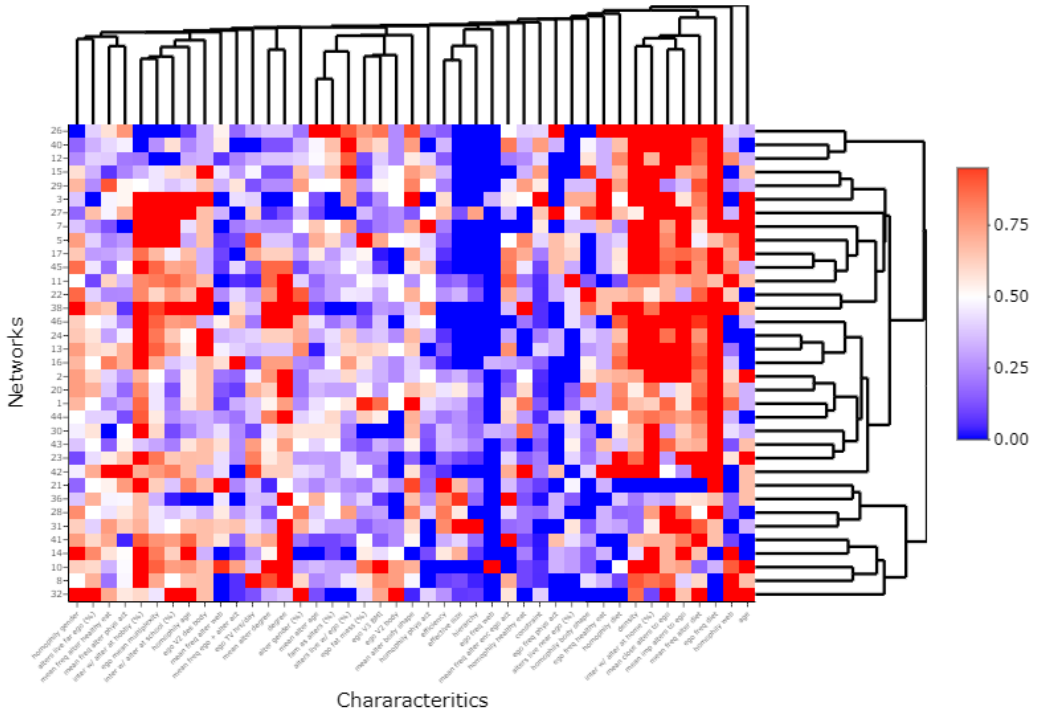

Figure 8: The heatmap of the data using the single linkage method to organize the placement of the network attribute data.

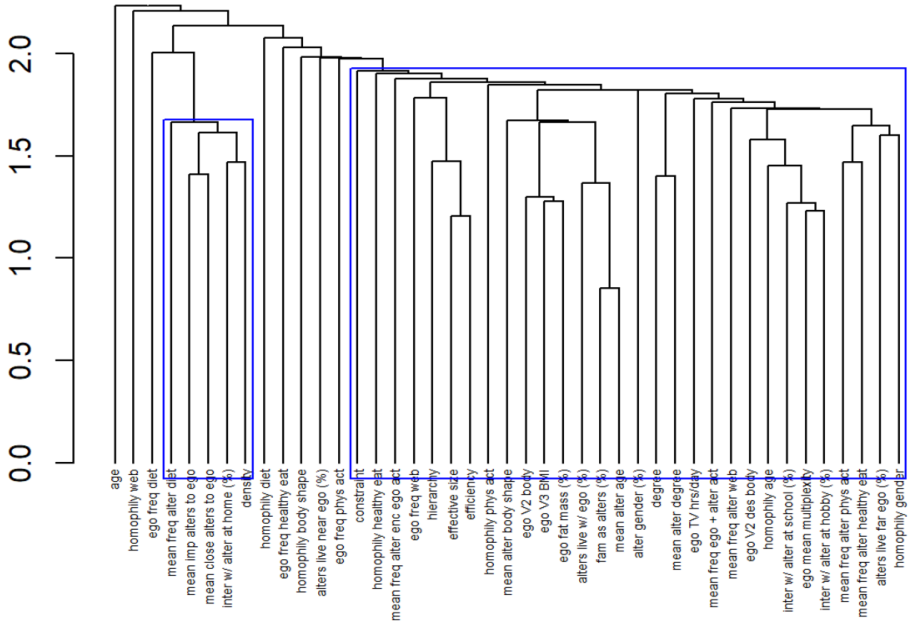

Figure 9: The dendrogram of the network attributes. The blue boxes indicate clusters with AUs of 95% or above. Notice that the clusters seem to form a chain from left to right.

Average Linkage Method:

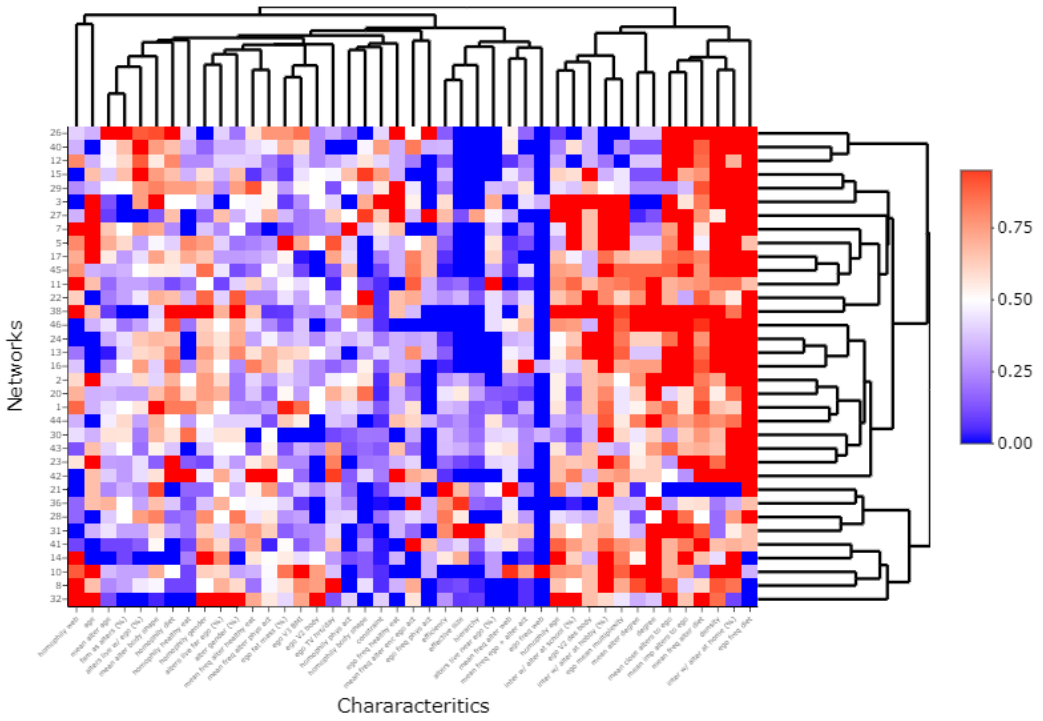

Figure 10: The heatmap of the data using the average linkage.

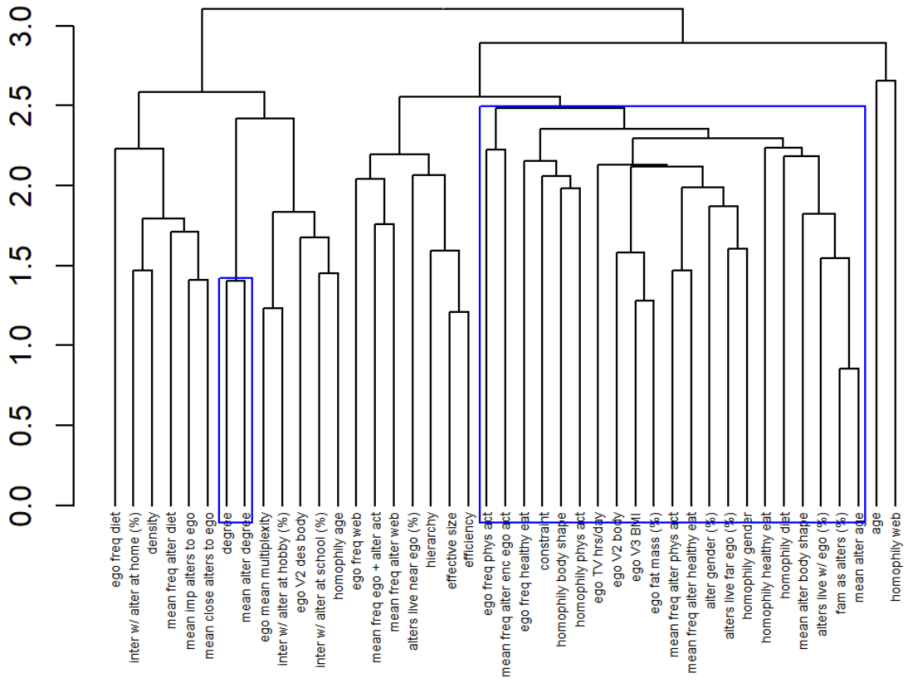

Figure 11: The dendrogram of the network attributes. The blue boxes indicate clusters with AUs of 95% or above. Notice that the significant clusters contain less elements than those of the single linkage method.

Complete Linkage Method:

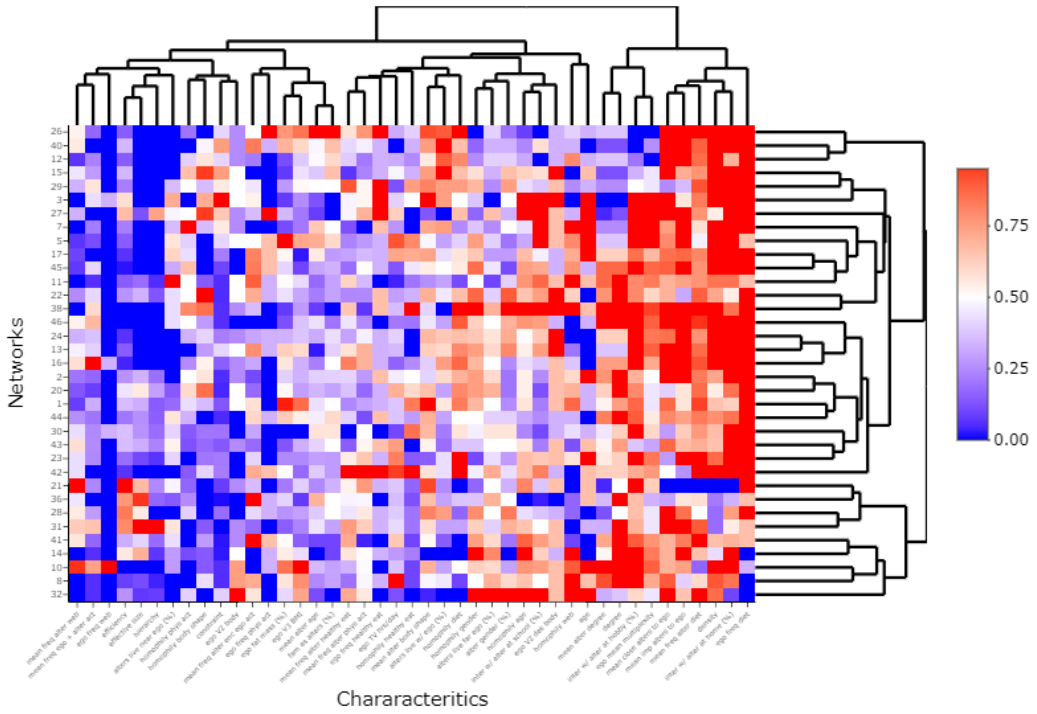

Figure 12: The heatmap of the data using the complete linkage method.

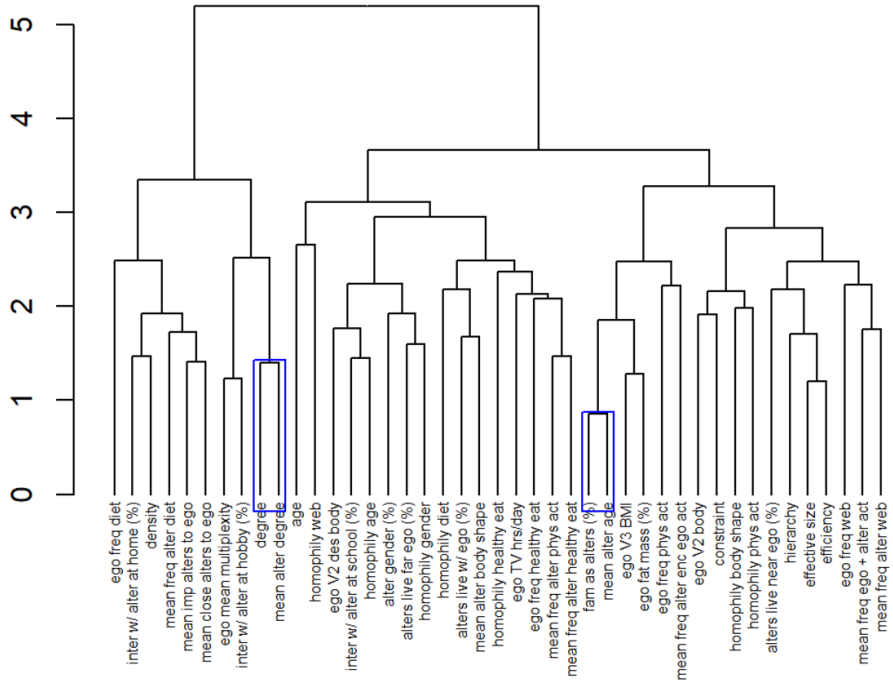

Figure 13: The dendrogram of the network attributes. The blue boxes indicate clusters with AUs of 95% or above. This method exhibits the smallest number of significant clusters.

Discussion

When comparing the node-link diagram (**fig. 7)** of the QUALITY social networks with the heatmaps **(fig. 8, 10, 12)**, the weaknesses of the former become obvious. The most notable shortcoming of the node-link diagram is the amount of information it can convey: for each graph, only 5 characteristics are described, whereas the heatmaps depicts 43. Another significant difference between the two kinds of visualization is the ease with which any number of networks can be compared. The organization of the networks in the node-link diagram makes it difficult to compare the various graphs; the heatmaps accomplish this task due to the ordering of the networks and variables, as well as the colour scheme. An additional distinction between the two visualization techniques is that when presented in the appropriate medium (i.e. in a dynamic report) the heatmaps are interactive, while the node-link diagram is static. The ability to focus on certain regions of the illustration increases its legibility. For these reasons, the heatmaps facilitated the exploration process of the QUALITY social network data.

Although the heatmaps outperform the node-link diagrams when it comes to data exploration, they do suffer from some limitations. Since they rely so heavily on hierarchical clustering to illustrate the patterns in the data, parameters changes used in the clustering (such as the similarity measure, the linkage method and the scaling technique), can affect the detected patterns. As with hierarchical clustering, many versions of a heatmap should be produced with many different methods in order to identify regular patterns in the data, as well as to determine which parameters yield the most reasonable results. For this reason, the analysis of the data was performed using the single, average and complete linkage methods **(fig. 8-13)**. In the future, additional similarity measures and linkage methods will be utilized to perform analyses. Unfortunately, regardless of the results, it is and will continue to be difficult to vouch for a method that has only been carried out on a single data set. Research aimed at further elucidating the advantages and disadvantages of using heatmaps as an instrument for multiple network data exploration is needed.

In fact, the most substantial issue hindering the development of this visualization technique is the lack of multi-network data with which to test it on. To further investigate its strengths and weaknesses, there are two possibilities. Either (1) additional multi-network data sets can be simulated, or (2) communities found in traditional network data sets can be split,

creating a pseudo-multi-network data. Testing the method on simulated data will no doubt offer insight on the optimal choice of hierarchical clustering methods to use when exploring data with a known set of characteristics. Furthermore, determining whether this approach is useful in the exploration of networks containing communities may extend the usefulness of this visualization method to other varieties of network data.

Conclusion

With the help of the new visualization technique, relationships among the graph, node and structural characteristics of the networks that are described within the childhood obesity and social network literature were captured. The heatmaps also illustrated the distribution of variables across networks and permitted the comparison of various elements of the data. However, since the analyses were performed on a single data set, it is impossible to conclude that heatmaps are an appropriate tool for the exploration and visualization of all multiple network data. Unfortunately, there is not an abundance of this kind of data; simulated networks will be used to further test and improve this method.

References

1. Hoaglin, D. C., Mosteller, F. & Tukey, J. W. *Understanding robust and exploratory data analysis*. (Wiley, 1983).
2. Perer, A. & Shneiderman, B. Integrating Statistics and Visualization for Exploratory Power: From Long-Term Case Studies to Design Guidelines. *IEEE Comput. Graph. Appl.* **29,** 39–51 (2009).
3. Newman, M. *Networks: An Introduction*. (Oxford Scholarship Online, 2010).
4. Demongeot, J., Hansen, O. & Taramasco, C. Discrete dynamics of contagious social diseases: Example of obesity. *Virulence* **7,** 129–140 (2016).
5. Gehlenborg, N. & Wong, B. Points of view: Networks. *Nat. Methods* **9,** 115–115 (2012).
6. Gehlenborg, N. & Wong, B. Points of view: Heat maps. *Nat. Methods* **9,** 213–213 (2012).
7. Lambert, M. *et al*. Cohort Profile: The Quebec Adipose and Lifestyle Investigation in Youth Cohort. *Int. J. Epidemiol.* **41,** 1533–1544 (2012).
8. James, G., Witten, D., Hastie, T. & Tibshirani, T. *An Introduction to Statistical Learning*. (Springer).
9. Empirical distribution. Available at: https://www.statlect.com/asymptotic-theory/empirical-distribution. (Accessed: 31st August 2017)
10. Paradis, E. *Analysis of phylogenetics and evolution with R*. (Springer, 2006).
11. Shimodaira, H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.* **32,** 2616–2641 (2004).
12. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).

13. Étude QUALITY ::: QUALITY Study. Available at: http://www.etudequalitystudy.ca/index.php?p=1&lang=en&mid=1. (Accessed: 5th September 2017)
14. Macdonald-Wallis, K., Jago, R., Page, A. S., Brockman, R. & Thompson, J. L. School-based friendship networks and children's physical activity: A spatial analytical approach. *Soc. Sci. Med. 1982* **73,** 6 (2011).
15. Daw, J., Margolis, R. & Verdery, A. M. Siblings, Friends, Course-mates, Club-Mates: How Adolescent Health Behavior Homophily Varies by Race, Class, Gender, and Health Status. *Soc. Sci. Med. 1982* **125,** 32–39 (2015).
16. de la Haye, K., Robins, G., Mohr, P. & Wilson, C. Obesity-related behaviors in adolescent friendship networks. *Soc. Netw.* **32,** 161–167 (2010).
17. Yu, J. Social Network Analysis of Egonets. (2017).