exeley

Article | DOI: 10.21307/connections-2018-002

CONNECTIONS

Issue 1 | Vol. 38

Techniques: Dichotomizing a Network

Stephen P. Borgatti[1]* and Eric Quintane[2]

[1]University of Kentucky, Gatton College of Business & Economics Lexington, KY.

[2]School of Management, The University of Los Andes, Bogota, Colombia.

*E-mail: sborgatti@uky.edu.

Abstract

This techniques guide provides a brief answer to the question: How to choose a dichotomization threshold? We propose a two step approach to selecting a dichotomization threshold. We illustrate the approaches using two datasets and provide instructions on how to perform these approaches in R and UCINET.

Keywords

Techniques, Dichotomization.

There are many reasons to dichotomize valued network data. It might be for methodological reasons, for example, in order to use a graph-theoretic concept such as a clique or an n-clan, or to use methods such as ERGMs or SAOMs, which largely assume binary data[1]. There is also the matter of visualizing networks, where fewer ties often yield a considerably more readable picture. It could also be for theoretical reasons. For example, in order to distinguish between positive and negative ties, since tie strength or valence is often captured using a single scale, which then needs to be dichotomized in order to match the theory. Finally, we might be engaging in a certain kind of data smoothing: we have collected data at fine levels of differences in the strength of tie, but are not confident that small differences are meaningful. We have greater confidence in a few big buckets such as strong and weak than in 100 graduations of strength.

Whatever the reason, if we are going dichotomize, the question is at what level should we dichotomize? In some cases, the situation is guided by theoretical meaningfulness and the research design. For example, suppose respondents are asked to rate others on a scale of 1 = do not know them, 2 = acquaintance, 3 = friend, and 4 = family. We see there is a loose gradation from "does not know" to "knows well"; however, categories 3 and 4 do not possess so much degrees of closeness as different kinds of social relations. The choice of which to use is determined by the research question. A similar example is provided by questions that ask for a range of effects from negative to positive. If respondents are asked to rate others on a scale of 1 = dislike a lot, 2 = dislike somewhat, 3 = neither like nor dislike, 4 = like somewhat, and 5 = like a lot, for many analyses, it will make sense to choose a cut off of >3 or >4 for positive ties and <3 or <2 for negative ties. Note that in both of the last examples, we are still confronted with a choice of two values to choose from. In addition, if the scale points are more ambiguous than the ones above, or if the data are counts or rankings, then there is likely no a priori way of deciding where to dichotomize.

Here, we propose a two-step approach to dichotomizing. Step 1 is to simply dichotomize at every level (or a collection of *k* bins) and examine the network produced at each level. Step 2 is to use simple analytics in order to obtain an informed rationale for a specific dichotomization threshold that makes sense for a given data set.

Step 1

For step 1, input your valued network into your favorite network data management software and dichotomize at every level of the scale (see insert for information about how to do this in R and in UCINET). We recommend always spending some time visualizing the

[1]There are, of course, many methods that do not require dichotomization. For example, we do not need to dichotomize in order to measure eigenvector centrality, nor to apply the relational event model (Butts, 2008).

Table 1. One mode DGG Women by Women network projection.

	EV	LA	TH	BR	CH	FR	EL	PE	RU	VE	MY	KA	SY	NO	HE	DO	OL	FL
EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0
THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0
FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0
ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0
PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1
RUTH	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1
VERNE	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1
MYRNA	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1
KATHERINE	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1
SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1
NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2
HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1
DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1
OLIVIA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2
FLORA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2

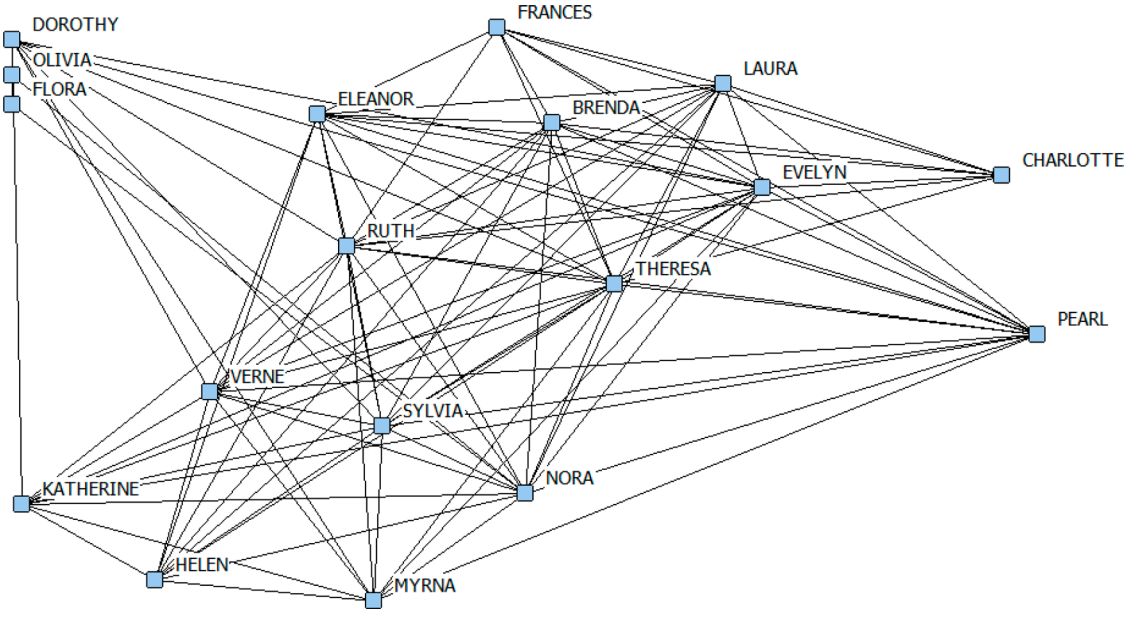

Figure 1: DGG Women by Women dataset dichotomized above 1.

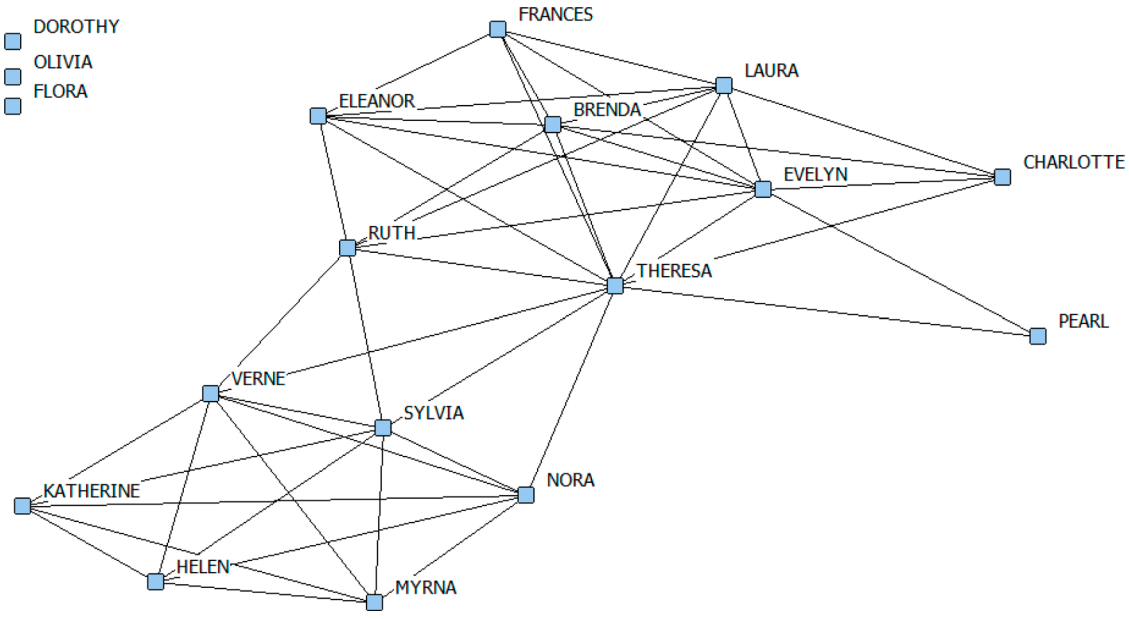

Figure 2: DGG Women by Women dataset dichotomized above 2.

networks, which can be very informative regarding the emergence of clusters at certain levels of dichotomization. For example, consider Davis et al.'s (1941) women-by-events data (often referred to as the Davis data set or the DGG data). We construct a 1-mode women-by-women network by multiplying the original by its transpose. The result is shown in Table 1.

If we dichotomize at >1 and visualize, we get Figure 1.

If we dichotomize at >2, we get Figure 2.

And if we dichotomize at >3, we get Figure 3.

Thus, the successive dichotomizations reveal a 2-group structure, which is illuminating[2]. In other networks, successive dichotomization confirms a core/periphery structure. For example, the BKFRAT data set (Bernard et al., 1980) gives the number of times each pair of actors was seen interacting by an observer. The values range from 0 to 51. If we dichotomize at > 0, we get Figure 4.

If we dichotomize at > 2, we get Figure 5.

Dichotomizing at > 4, we get Figure 6.

Dichotomizing at > 6, we get Figure 7.

And so on. Core-periphery structures have a kind of self-similarity property where the main component always looks the same regardless of what level of dichotomization produced it.

Step 2 (three approaches)

Now, successive dichotomizations are informative, but our original question was about choosing a single dichotomization that would be used in all further analyses, which is where step 2 becomes important. For step 2, we present three potential approaches. The first will horrify some people. This approach is to choose the level of dichotomization that maximizes your results. For example, suppose you are predicting managers' performance as a function of betweenness centrality. For each possible level of dichotomization, you measure betweenness centrality and regress performance on betweenness, along with any control variables. The level of dichotomization that yields the highest r^2 is the one you choose.

As we said, some people (scientists, statisticians, and people of good character) will be horrified[3]. There is definitely a danger of overfitting. The predictions work really well for this one data set, but perhaps not

[2]However, this should not be taken as definitive. Various normalizations of the data, as well as bipartite representations, tend to show a third smaller subgroup. See Freeman (2003) for a related discussion.

[3]On the other hand, these same people are happy to use regression to find the optimal coefficients to show a relationship between their explanatory variable and a dependent. Perhaps, we should ask them to choose the coefficients *a priori* on the basis of strong theory.

[4]Of course, if you have these other datasets on hand, then you could pick the level of dichotomization that yields the highest average r^2 across all datasets. The same applies if you have multiple DVs and IVs – you pick that level of dichotomization that gives the best results across all datasets, DVs, and IVs.

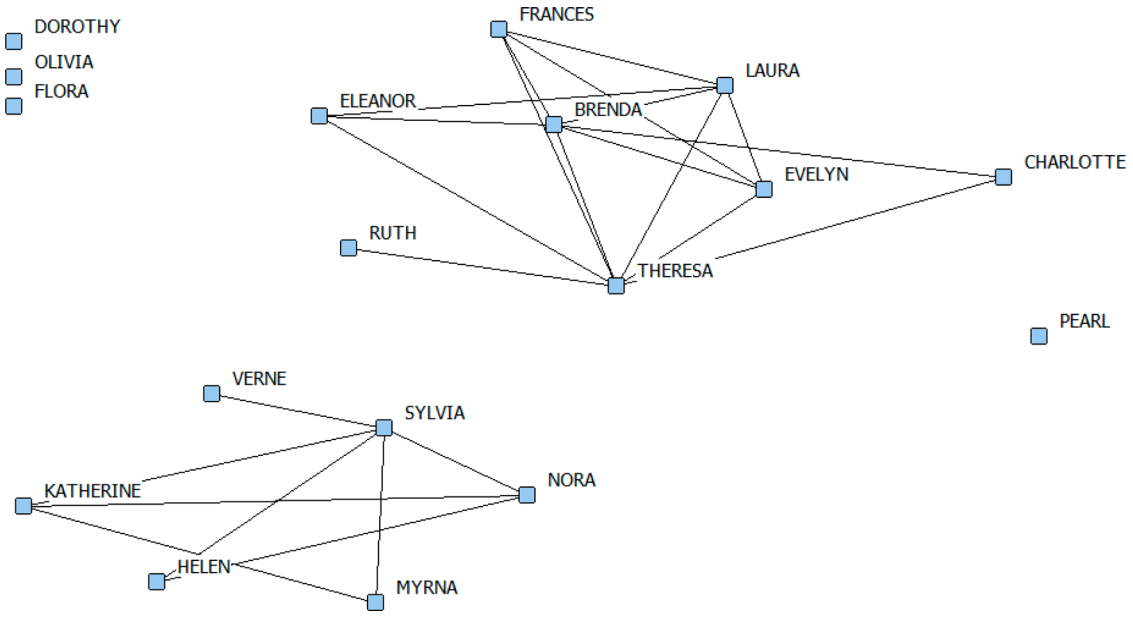

Figure 3: DGG Women by Women dataset dichotomized above 3.

for others[4]. The other issue is that the particular dichotomization value that scores highest may be an odd value that you cannot explain. For example, suppose we carry out this procedure and get the results shown in Table 2.

Clearly, we would choose 5, but how to make sense of these results? They rise and fall with no rhyme or reason. In this case, we would strongly advise against taking this approach. On the other hand, if the results were something like those presented in Table 3, we would be comforted by the underlying regularity and feel good about choosing 5, even though we might be hard-pressed to explain why medium density worked best.

A slightly less controversial version of this approach might be to choose the dichotomized version of your network that maximizes the replication of results from past studies. For example, we know from past studies that actors with higher levels of self-monitoring are more likely to receive

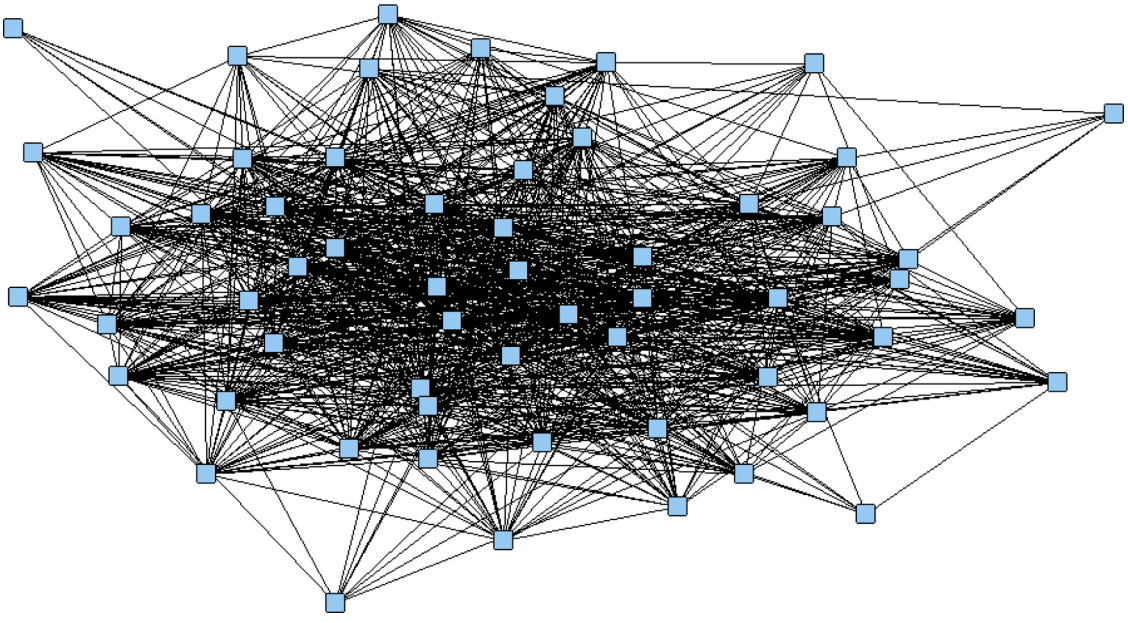

Figure 4: BKS FRATERNITY dataset dichotomized above 0.

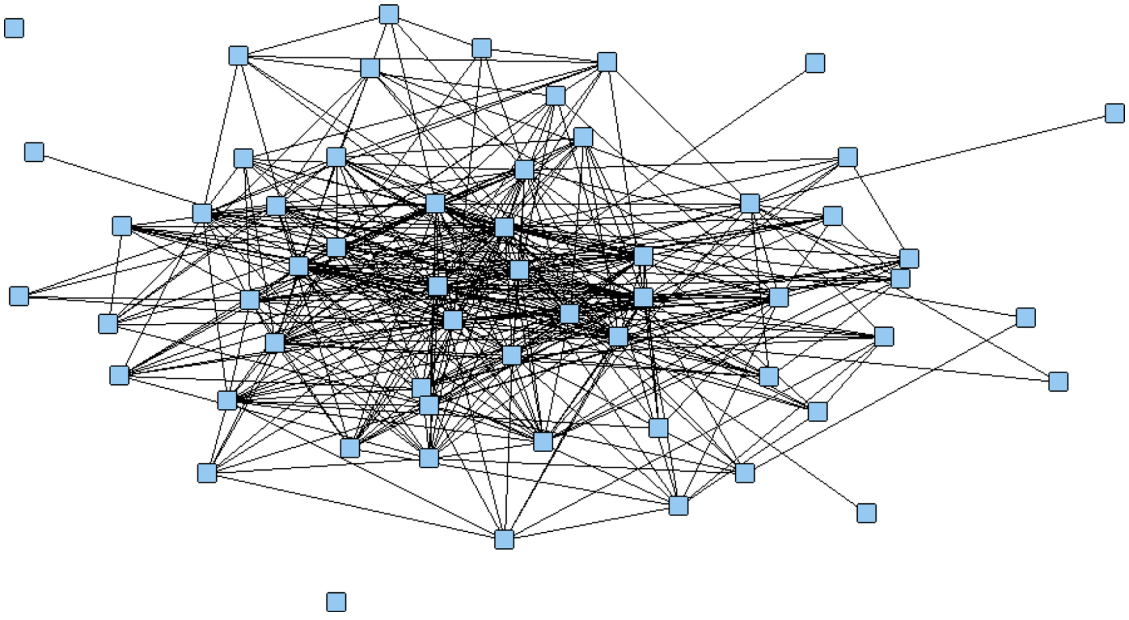

Figure 5: BKS FRATERNITY dataset dichotomized above 2.

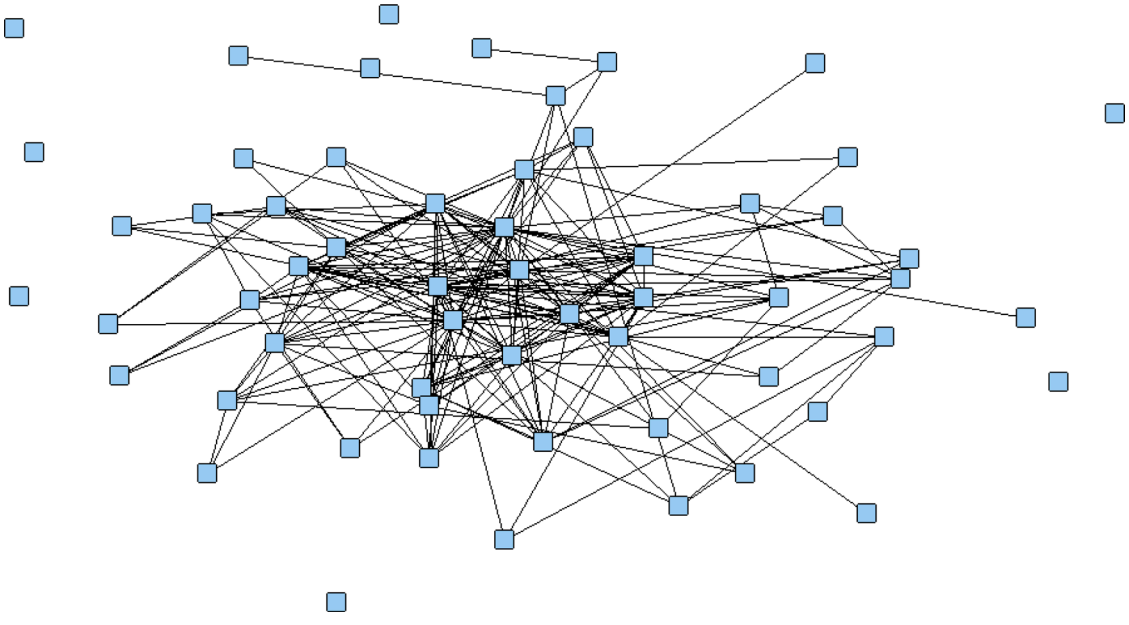

Figure 6: BKS FRATERNITY dataset dichotomized above 4.

more friendship nominations. We could choose the dichotomization threshold that maximizes the relationship between self-monitoring and new friendship nominations, even if the test of our hypothesis has to do with betweenness centrality and performance.

That was the first approach. The second approach is less controversial. Dichotomization, by its very nature, is a distortion of the data[5]. Where once you had nuance, you now have just 'has tie' and 'not tie.' This does violence to your data. The question is, how

[5]Clearly, in some cases, distorting the data is what we are looking for, for example, when distinguishing between negative and positive ties. In this case, we should not expect the dichotomized data to preserve the properties of the original dataset and we should either use a theoretically or literature driven approach or revert to approach 1.

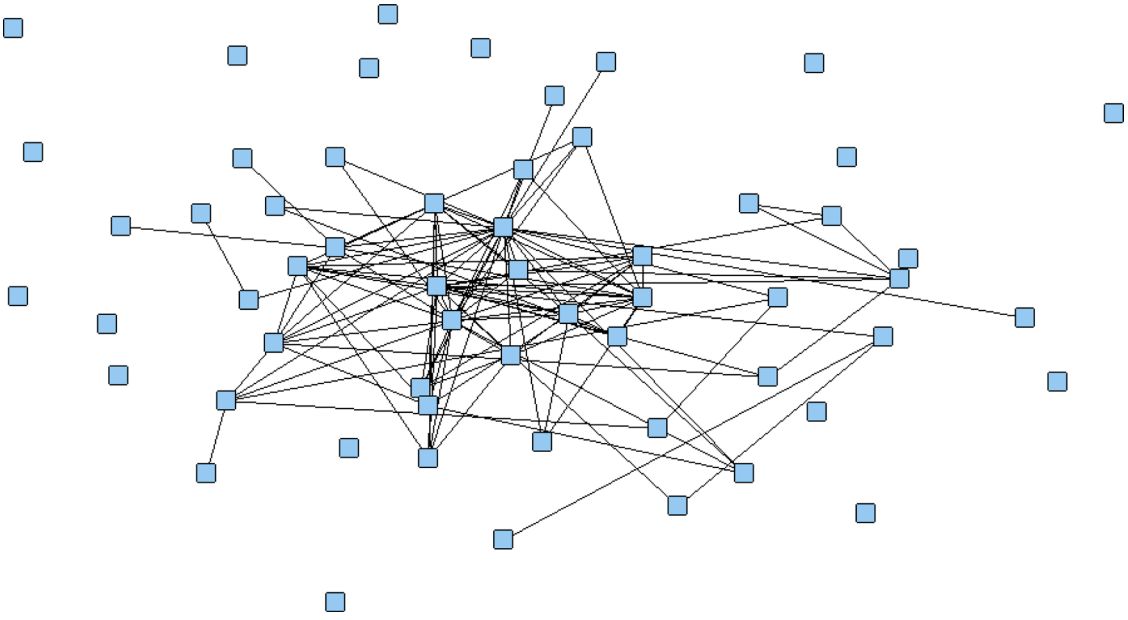

Figure 7: BKS FRATERNITY dataset dichotomized above 6.

much? Suppose, as in an analysis of variance, you predicted your valued data from your dichotomized data. Some cutoff values are going to yield better predictions than others. Here is an example using the Davis, Gardner, and Gardner women-by-women data. In the table below, the first column is the dichotomization value. For example, value 4 means that the data were dichotomized at ≥ 4. Dichotomizing at ≥ 4 results in a network with 48 ties, which corresponds to a density of 0.16. The interesting part is the correlation column, which achieves its maximum at ≥ 3 (correlation 0.81). The correlation refers to the correlation between the original valued matrix and the dichotomized matrix. A correlation of 0.81 is extremely high. Yes, the data

Table 2. R-square of models predicting performance using betweenness centrality at different levels of dichotomization.

Dichot. level	R^2
1	0.05
2	0.29
3	0.02
4	0.01
5	0.31
6	0.06
7	0.11
8	0.02
9	0.23

Table 3. R-square of models predicting performance using betweenness centrality at different levels of dichotomization.

Dichot. level	R^2
1	0.05
2	0.09
3	0.12
4	0.23
5	0.31
6	0.27
7	0.22
8	0.15
9	0.07

Table 4. Z-score, correlation, number of ties and density of the DGG dataset at different dichotomization levels.

Value	*Z*-score	Correlation	Ties	Density
7	3.352	0.271887	2	0.006536
6	2.667	0.646625	16	0.052288
5	1.983	0.666829	18	0.058824
4	1.298	0.781314	48	0.156863
3	0.613	0.811928	92	0.300654
2	–0.072	0.720115	190	0.620915
1	–0.756	0.457341	278	0.908497
0	–1.441		306	1.000000

are distorted by dichotomizing, but the dichotomized matrix still retains a very high resemblance to the original data. We have chosen a level of dichotomization that does the least violence to the original data (Table 4).

Interestingly, ≥ 3 is the level just below the one at which the network splits into two large components (along with four isolates). At ≥ 4, the network looks like this, as shown in Figure 8

The third approach is theory based, and can be harder to implement. There are certain cases where we can use the emergent properties of the dichotomized networks themselves in order to identify the correct dichotomization threshold, just like when we noticed the appearance of clusters while visually inspecting different dichotomization thresholds in the DGG data. As an example, let us consider an approach proposed by Freeman (2003) to distinguish between weak and strong ties. In his piece on the strength of weak ties, Granovetter (1973) argues that an important characteristic of strong ties is that if A is strongly tied to B, and B is strongly tied to C, then A is likely to be at least weakly tied to C. In his analysis of the DGG data, Freeman (2003) refers to Granovetter's transitivity rule as g-transitivity. A data set is perfectly g-transitive if there are no violations of g-transitivity. Given a valued data set (and selecting a value such as zero as an indicator of no ties), Freeman's proposal is to dichotomize the data set at every possible cutoff and calculate the number of violations of g-transitivity at each level. The lowest cutoff with an acceptable number of violations (such as zero) identifies the strong tie. For example, applied to the Davis women data, we get Table 5.

Table 5. Number of g-transitive and intransitive triples in the DGG dataset at different dichotomization levels.

Value	Trans	Intrans
7	0	0
6	26	0
5	30	0
4	160	0
3	526	4
2	2,032	44
1	3,786	292
0	4,448	448

The table shows that at ≥ 4, the number of g-transitive triples is 160 and the number of intransitive triples is 0. Hence, ties 4 or above are strong ties, and ties < 4 but > 0 are weak ties.

Combining this with our previous approach, we might summarize the situation as follows. Dichotomizing at ≥ 3 optimally identifies ties of any kind in terms of the least-violence criterion, and maintains a single large component (plus isolates). Dichotomizing at ≥ 4 identifies strong ties, which strongly fragment the network. The latter is useful for sharply outlining a subgroup structure, while the former enables the calculation of measure that requires connected networks (aside from isolates) (Figure 9).

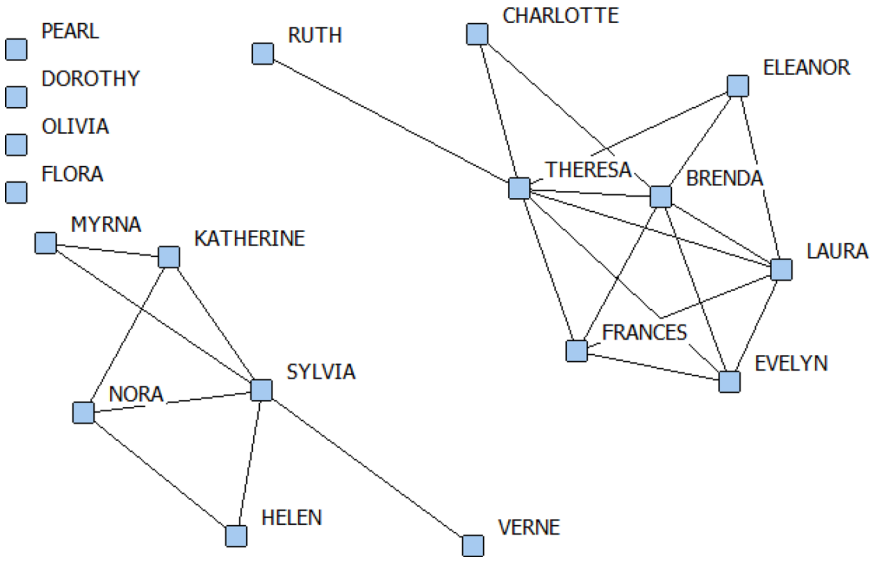

Figure 8: DGG Women by Women dataset dichotomized at 4.

It is worth noting that Freeman's approach needs not be limited to maximizing g-transitivity. On theoretical grounds, we may identify a specific mechanism that organizes ties. For example, we may see a status mechanism such at the Matthew effect in which nodes that already have a lot of ties tend to attract even more ties. Now, to dichotomize valued data, we choose the cutoff that maximizes the extent to which there are just a few nodes with many ties and a great many nodes with few ties. Alternatively, we might choose the cutoff to maximize the level of transitivity in the network.

Conclusion

This "How to" guide on dichotomization is intended to provide guidance on how to find a suitable threshold for dichotomization for social network data. We propose that in all cases, we should start by creating multiple versions of the dichotomized network at every possible value of the threshold and inspect them visually. Then, we suggest three separate approaches in order to choose (and justify your choice of) a single threshold based on (i) maximiz-

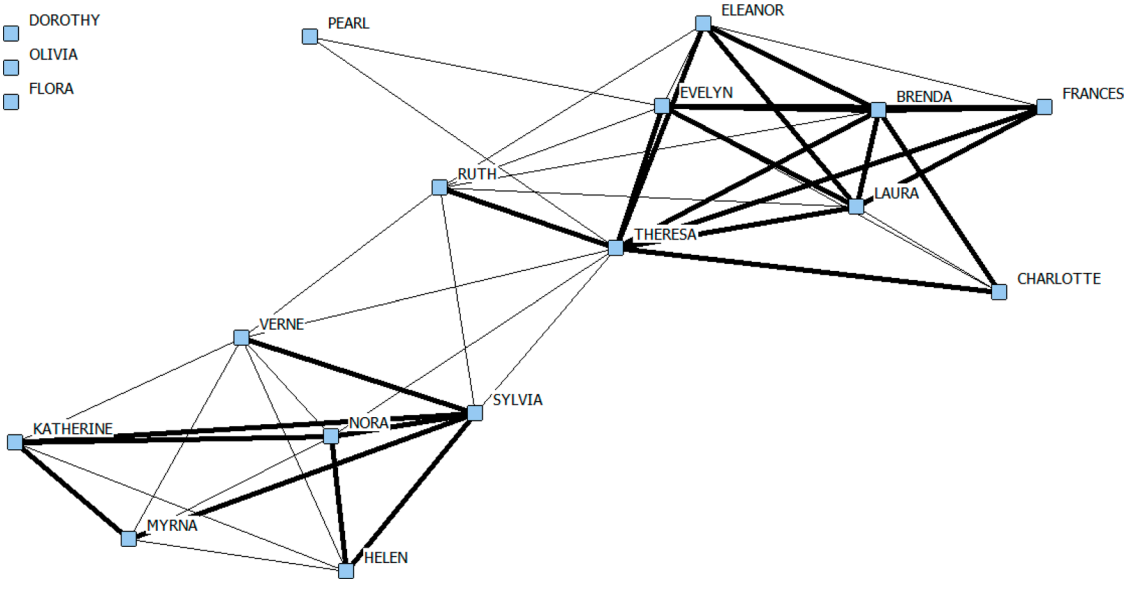

Figure 9: DGG Women by Women dataset dichotomized at 3. Strong ties in bold.

ing expected results, (ii) minimizing distortions, and (iii) identifying specific emergent properties in the network.

References

Bernard, H., Killworth, P. and Sailer, L. 1980. Informant accuracy in social network data IV. *Social Networks* 2: 191–218.

Butts, C.T. 2008. A relational event framework for social action. *Sociological Methodology* 38: 155–200.

Davis, A., Gardner, B. B. and M. R. Gardner 1941. Deep South, Chicago: The University of Chicago Press.

Freeman, L.C. 2003. Finding social groups: a meta-analysis of the southern women data, in Breiger, R., Carley, K., and Pattison, P. (Eds), *Dynamic social network modeling and analysis: workshop summary and papers* Committee on Human Factors, National Research Council: 39–45, National Academies Press.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 81: 1287–303.

Addendum 1 – R Script

```
#Import the Davis data set in R, assuming that it is already in a text file, for example exported from UCINET.
    library(readr)
    davis <- as.matrix(read.csv("davis.txt",sep = "\t", row.names = 1))
    #Create a one-mode network by multiplying the original matrix by its transpose
    davisonemode <- davis %*% t(davis)
    diag(davisonemode) <- 0
    #Dichotomize the network at all values
    davisonemodedic <- array(dim = c(NROW(davisonemode),NCOL(davisonemode),max(davisonemode)))
    for (i in 1:max(davisonemode)) {
    davisonemodedic[,,i] <- ifelse(davisonemode>=i, 1, 0)
    }
    #Visualize all networks
    library(sna)
    par(mfrow = c(4,2))
    for (i in 1:max(davisonemode)) {
    plot(as.network(davisonemodedic[,,i]))
    }
    #Correlation between original network and dichotomized networks, and some descriptive statistics
    stats <- array(dim = c(max(davisonemode),4))
    colnames(stats) <- c("Threshold", "Correlation", "Num of 1 s", "Density")
    for (i in 1:max(davisonemode)) {
    stats[i,1] <- i
    stats[i,2] <- summary(qaptest(list(davisonemode, davisonemodedic[,,i]), gcor, g1 = 1, g2 = 2))$test
    stats[i,3] <- sum(davisonemodedic[,,i])
    stats[i,4] <- stats[i,3]/(NROW(davisonemode)*(NROW(davisonemode) - 1))
    }
    stats
```

Addendum 2 – UCINET

To visualize successive dichotomizations in UCINET, one opens the valued data as usual and presses the + sign in the rels tab at right to raise the level of dichotomization by one unit, see Figure A1, below.

This can also be done in the command line interface (CLI) as follows:

```
->d1 = dichot(women ge 1)
->d2 = dichot(women ge 2)
->d3 = dichot(women ge 3)
Etc.
```

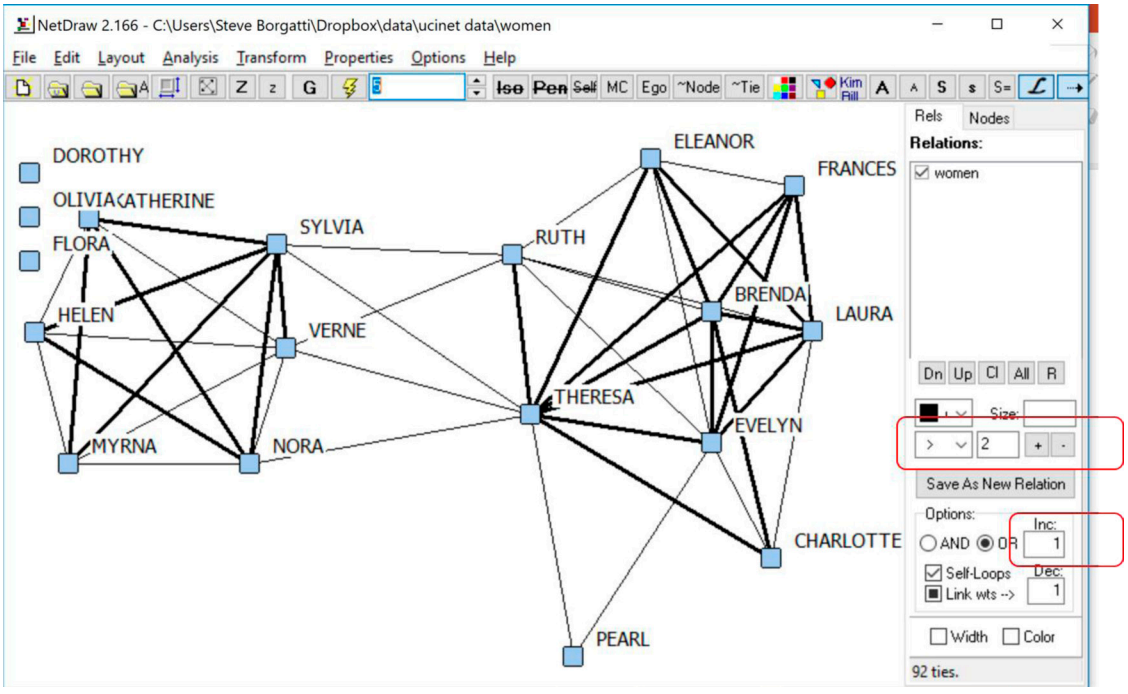

Figure A1: Screenshot of Netdraw.

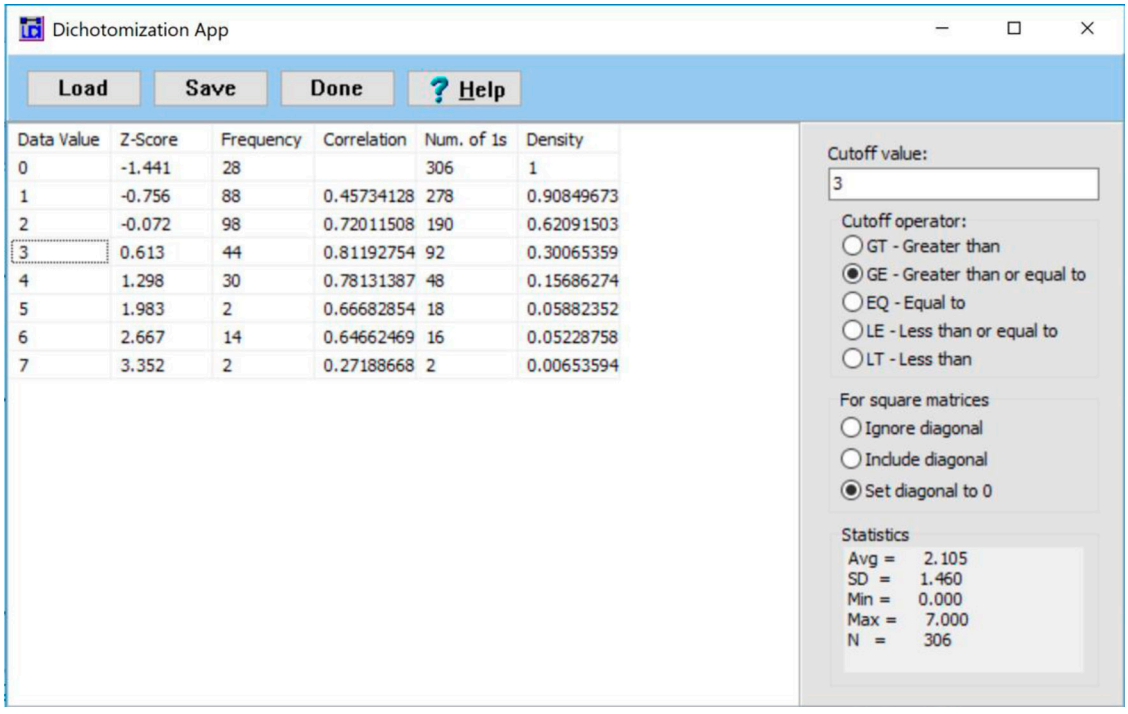

Figure A2: Screenshot of UCINET's Interactive Dichotomization routine's results.

In addition, the network could be drawn after each step:

->draw d1
->draw d2
Etc.

To compute the correlation between an original data set and successive dichotomizations of it, we can use UCINET's Transform|Interactively Dichotomize procedure. Figure A2 below shows this procedure applied to the DGG women data.

Finally, to execute Freeman's strong-weak-null tie decomposition based on g-transitivity, we can use UCINET's command line interface (CLI) as shown in Table A1.

Table A1. G-transitivity decomposition command line instruction and output in UCINET.

->dsp gtrans(women)

Level	1 Trans	2 Intrans	3 Possible	4 Prop Trans
7	0	0	0	
6	26	0	26	1
5	30	0	30	1
4	160	0	160	1
3	526	4	530	0.992
2	2,032	44	2,076	0.979
1	3,786	292	4,078	0.928
0	4,448	448	4,896	0.908