Metrics Homework

Ben Zuckier

Network Science

1 Getting Started

a)

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The above is the adjacency matrix for our network of 12 vertices, with names: Asher, Benjamin, Dan, Gad, Issachar, Judah, Levi, Naphtali, Reuben, Simeon, Yosef, Zebulun. I will sometimes refer to them as: A, B, D, G, I, J, L, N, R, S, Y, Z. Note that they are sorted Lexicographically (all unique letters) and their respective ordering is their vertex number in the network (A is 1, B is 2, D is 3, ..., Z is 12). This is how they have been arranged in the matrix above.

b) The Perron-Frobenius theorem states that the largest eigenvalue is unique and its corresponding eigenvector has only positive entries.

Mathematica Session

In[1]:= **eigvals = N[Sort[Eigenvalues[adjMatrix], Greater], 2]**

Out[1]= $\{3.3, 2.0, 1.4, 0.91, 0.42, 0.21, -0.33, -0.45, -1.0, -1.6, -2.1, -2.8\}$

In[2]:= **bigEigvalTally = Tally[eigvals][[1]]**

Out[2]= $\{3.3, 1\}$

In[3]:= **bigEigvec = N[Eigenvectors[adjMatrix], 2]**

Out[3]= $\{0.90, 1.9, 1.2, 0.27, 0.91, 1.3, 0.28, 0.94, 0.95, 1.2, 0.85, 1.0\}$

In[4]:= **Length[Select[bigEigvec, # ≤ 0&]]**

Out[4]= 0

In the above, we get the eigenvalues for our adjacency matrix and in Out[2] we see that the greatest eigenvalue, equal to about 3.3, is unique (occurs just once) – via the `Tally` operator. After getting the corresponding eigenvector, it is shown in Out[4] that there are no entries that are less than zero, i.e. all entries all positive.

Hence, the largest eigenvalue is unique and its corresponding eigenvector has all positive entries. Thus we have shown that our adjacency matrix's eigensystem satisfies the theorem.

2 Most Important Student?

Each of these metrics is listed for nodes A, B, D, G, I, J, L, N, R, S, Y, Z in order.

a) Degree centrality:

$\{3, 6, 3, 1, 4, 3, 1, 3, 2, 3, 2, 3\}$

b) Eigenvector centrality:

$\{.0773, .165, .0985, .0235, .078, .107, .0241, .0801, .0816, .106, .0731, .0855\}$

Note that we have used the normalized eigenvector centrality. For the non-normalized version, the eigenvector corresponding to the largest eigenvalue (which is the non-normalized eigenvector centrality) can be found in Out[3] of the Mathematica section of part (b) above.

c) Katz centrality:

Here I calculated the Katz (normalized) centrality for values of $\alpha = \{.301009, 0.301008, 0.301007, 0.301, 0.3, 0.29, 0.2, 0.15, 0.1, 0.0001\}$ and $\beta = 1$. The first value of alpha in this list, 0.301009, is slightly greater than the inverse of the largest eigenvalue – $\frac{1}{\kappa_1}$ – which is the value at which the centrality begins to diverge. The values for α where the centrality converges and provides meaningful information lie in the range $[0, \kappa_1^{-1})$. For part (g) below, we will take the advice of Newman (Newman pp. 164) and use a value for alpha very close to κ_1^{-1} which yields a centrality very similar to the (normalized) Eigenvector centrality above.

$\alpha = 0.301009:$

$$\textbf{Normalize}[\{-392773., -837149., -500998., -119426., -396756., -545650., -122540., -407102., -414868., -541113., -371416., -434477.\}]$$

$\alpha = .301008:$

$$\{.0772526, .164655, .0985388, .0234895, .0780362, .107321, .024102, .080071, .0815984, .106429, .073052, .0854551\}$$

$\alpha = .301007:$

$$\{.0772527, .164654, .0985387, .0234899, .0780364, .107321, .0241023, .0800711, .0815982, .106428, .073052, .0854551\}$$

$\alpha = .301$:

$$\{.0772537, .164651, .0985382, .0234921, .078038, .107319,\\ .0241042, .0800715, .0815973, .106427, .0730522, .0854553\}$$

$\alpha = .3$:

$$\{.0773887, .164261, .0984686, .0238073, .0782582, .107105,\\ .0243706, .0801358, .0814615, .106178, .0730855, .0854796\}$$

$\alpha = .29$:

$$\{.0786256, .160482, .0977626, .0268596, .080298, .105078,\\ .0269881, .0807409, .0802301, .103852, .0733993, .0856835\}$$

$\alpha = .2$:

$$\{.083941, .133102, .0914833, .0488701, .0897349, .0929541,\\ .0476973, .0838712, .0758026, .0912949, .0754906, .0857576\}$$

$\alpha = .15$:

$$\{.0845565, .120412, .0885347, .0586168, .0907673, .0890294,\\ .0576612, .0843962, .0762417, .0878542, .0766786, .0852515\}$$

$\alpha = .1$:

$$\{.0844571, .108264, .0861524, .0674045, .0898895, .0862717,\\ .0668529, .084373, .077807, .0856499, .078231, .084647\}$$

$\alpha = .0001$:

$$\{.0833347, .0833597, .0833347, .0833181, .0833431, .0833347,\\ .0833181, .0833347, .0833264, .0833347, .0833264, .0833347\}$$

Here we see that as $\alpha \to 0$, the centrality of every node approaches $\frac{1}{n} = \frac{1}{12} = .08\overline{3}$.

The pairs of alpha from above for which the centralities have the same relative ranking (i.e. `Order`) are the following:
{0.301008, 0.301007}, {0.301008, 0.301}, {0.301008, 0.3}, {0.301007, 0.301}, {0.301007, 0.3}, {0.301, 0.3}, {0.15, 0.1}, {0.15, 0.0001}, {0.1, 0.0001}

d) PageRank centrality:

For PageRank (normalized), we again have two parameters to set. I have set $\beta = \frac{1}{n} = \frac{1}{12}$ and we are using values for $\alpha \in [0, 1]$ since we have an undirected network (Newman pp. 166). For part (g) below, I will be using $\alpha = .85$ which is the value Google uses.

$\alpha = 1$:

$$\{0.08824, 0.1765, 0.08824, 0.02941, 0.1176, 0.08824, 0.02941,\\ 0.08824, 0.05882, 0.08824, 0.05882, 0.08824\}$$

$\alpha = 0.9$:

$$\{0.08874, 0.1664, 0.08586, 0.03528, 0.1198, 0.0848, 0.03604,\\ 0.09235, 0.05886, 0.08522, 0.06024, 0.08648\}$$

$\alpha = 0.85$:

$$\{0.08861, 0.1618, 0.08482, 0.03803, 0.1201, 0.0836, 0.03905,\\ 0.09372, 0.0593, 0.08431, 0.06094, 0.08574\}$$

$\alpha = 0.8$:

$$\{0.08833, 0.1573, 0.08389, 0.04069, 0.1201, 0.08264, 0.04193, 0.09474, 0.05995, 0.08365, 0.06167, 0.0851\}$$

$\alpha = 0.7$:

$$\{0.08747, 0.1486, 0.08236, 0.04587, 0.1193, 0.08129, 0.04738, 0.09591, 0.06168, 0.0829, 0.06321, 0.08406\}$$

$\alpha = 0.5$:

$$\{0.08535, 0.1315, 0.08065, 0.05597, 0.1144, 0.08019, 0.05761, 0.09566, 0.06639, 0.08259, 0.06693, 0.08277\}$$

$\alpha = 0.1$:

$$\{0.08295, 0.09406, 0.08209, 0.0773, 0.0922, 0.0821, 0.0779, 0.08712, 0.07934, 0.08327, 0.07887, 0.08278\}$$

$\alpha = 0.01$:

$$\{0.08327, 0.08444, 0.0832, 0.08271, 0.0843, 0.0832, 0.08278, 0.08375, 0.08292, 0.08333, 0.08285, 0.08327\}$$

$\alpha = 0$:

$$\{0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333, 0.08333\}$$

Again, as $\alpha \to 0$, the centralities of each vertex converges to $\frac{1}{n} = \frac{1}{12} = .08\overline{3}$

The pairs of alpha from above which produce identical orderings are:
{0.9, 0.85}, {0.1, 0.01}, {0.9, 0.8}, {0.85, 0.8}, {0.7, 0.5}.

It seems like the centrality is more sensitive to alpha for values near the extremes of the range.

e) Closeness centrality:

The closeness centrality for vertex i is given by $C_i = \dfrac{1}{l_i} = \dfrac{n}{\sum_j d_{ij}}$ where l_i is the mean shortest distance from i to every node, and d_{ij} is the shortest distance from i to j.

$\mathbf{C} = \{.5455, .7059, .5714, .3871, .5714, .5455, .3871, .5714, .4615, .5, .5714, .5217\}$
$\mathbf{l} = \{1.833, 1.417, 1.75, 2.583, 1.75, 1.833, 2.583, 1.75, 2.167, 2., 1.75, 1.917\}$

A footnote on Newman page 171 explains that sometimes the distance from i to itself is not includeds so that $l_i = \dfrac{\sum_{j(\neq i)} d_{ij}}{n-1}$, with C_i still as its inverse. This is the standard implementation of the closeness centrality in Wolfram's Mathematica. Values here:

$\mathcal{C} = \{.5, .6471, .5238, .3548, .5238, .5, .3548, .5238, .4231, .4583, .5238, .4783\}$
$\ell = \{2., 1.545, 1.909, 2.818, 1.909, 2., 2.818, 1.909, 2.364, 2.182, 1.909, 2.091\}$

Newman also has formula 7.22 on page 172 that defines closeness based on the harmonic distance as $C_i' = \dfrac{1}{n-1} \displaystyle\sum_{j(\neq i)} \dfrac{1}{d_{ij}}$. He (bitterly) notes that "Despite its desirable qualities, however, Eq. (7.22) is rarely used in practice. The author has seen it employed only occasionally." To make him happy, I have calculated it here:

$\mathbf{C} = \{.5909, .7576, .6061, .4167, .6364, .5909, .4091, .6061, .5076, .5682, .5758, .5833\}$

For (g), we will be using the first definition, **C** and **l**.

f) Betweenness centrality:

For betweenness centrality we again have multiple options. Equation 7.24 on page 175 of Newman defines the centrality for node i as

$$x_i = \sum_{st} \frac{n^i_{st}}{g_{st}}$$

where n^i_{st} is the number of shortest paths from s to t that pass through i, and g_{st} is the total number of shortest paths from s to t. Note that for all of these equations, if n^i_{st} or g_{st} or both are equal to zero, then the output is zero.

On page 176 he gives Equation 7.25 which is a normalization of equation 7.24 defined as

$$\text{Normalized}[x_i] = \frac{1}{n^2} \sum_{st} \frac{n^i_{st}}{g_{st}} = \frac{x_i}{n^2}$$

where n is the number of vertices in our graph.

In a footnote on the same page, he gives Freeman's normalization as

$$\mathcal{N}[x_i] = \frac{1}{n^2 - n + 1} \sum_{st} \frac{n^i_{st}}{g_{st}} = \frac{x_i}{n^2 - n + 1}$$

Wikipedia's formula differs in that it only considers $s - t$ paths where $s \neq i \neq t$. This yields:

$$\mathcal{X}_i = \sum_{st|s\neq i\neq t} \frac{n^i_{st}}{g_{st}}$$

Wikipedia's normalization scheme normalizes $\mathcal{X}$ between 0 and 1 by scaling it, thereby preserving the relative ratios. It is done as follows:

$$\text{Normalized}[\mathcal{X}_i] = \frac{\mathcal{X}_i - \text{Min}[\mathcal{X}]}{\text{Max}[\mathcal{X}] - \text{Min}[\mathcal{X}]}$$

The standard Wolfram `BetweennessCentrality` is not normalized and gives exactly half the value of the Wikipedia implementation $\mathcal{X}$. My guess is that they are counting the *set* of the $s - t$ paths where $s \neq i \neq t$, ie no duplicates (for example it does not consider both the path from 1 to 2 and 2 to 1), so we have half the number of total sums, and hence half the resulting value. Thus:

$$\chi_i = \sum_{\{st|s\neq i\neq t\}} \frac{n^i_{st}}{g_{st}} = \frac{\mathcal{X}_i}{2}$$

.

Here are the final outputs for each.

$$\begin{aligned}
\mathbf{x} &= \{36., 68.33, 30.67, 22., 48.5, 29.33, 22., 45.5, 22., 23.67, 32., 32.\} \\
\text{Normalized}[\mathbf{x}] &= \{.25, .475, .213, .153, .337, .204, .153, .316, .153, .164, .222, .222\} \\
\mathcal{N}[\mathbf{x}] &= \{.271, .514, .231, .165, .365, .221, .165, .342, .165, .178, .241, .241\} \\
\mathcal{X} &= \{14., 46.3, 8.67, 0., 26.5, 7.33, 0., 23.5, 0., 1.67, 10., 10.\} \\
\text{Normalized}[\mathcal{X}] &= \{.302, 1., .187, 0, .572, .158, 0, .507, 0, .036, .216, .216\} \\
\chi &= \{7., 23.2, 4.33, 0., 13.3, 3.67, 0., 11.8, 0., 0.833, 5., 5.\}
\end{aligned}$$

Interesting, the relative ordering between all of these is the same!

For part (g) I will be using Newman's normalization, Normalized[**x**].

g) All centralities:

	Degree	Eigenvector	Katz	PageRank	Closeness	Betweenness
Asher	3	.0773	.0773	.0886	.5455	.25
Benjamin	6	.165	.1647	.1618	.7059	.475
Dan	3	.0985	.0985	.0848	.5714	.213
Gad	1	.0235	.0235	.038	.3871	.153
Issachar	4	.078	.078	.1201	.5714	.337
Judah	3	.107	.1073	.0836	.5455	.204
Levi	1	.0241	.0241	.0391	.3871	.153
Naphtali	3	.0801	.08	.0937	.5714	.316
Reuben	2	.0816	.0816	.0593	.4615	.153
Simeon	3	.106	.1064	.0843	.5	.164
Yosef	2	.0731	.0731	.0609	.5714	.222
Zebulun	3	.0855	.0855	.0857	.5217	.222
Spread	6	7.01	7.01	4.254	1.824	3.106

h) Spread:

	Degree	Eigenvector	Katz	PageRank	Closeness	Betweenness
Spread	6	7.01	7.01	4.254	1.824	3.106

The fact that the eigenvector/Katz centrality have the highest spread makes sense to me. The degree centrality for a vertex is already going to have a large spread since different vertices in our network have vastly different numbers of connections ie degree (one vertex only has 1, while another has 6). Not only does eigenvector/Katz take that into account, but it also scales it higher if you have important connections, which in our network means having popular friends. And it is usually the case that popular/sociable people are themselves friends with people who are popular. As such these metrics have a very large spread in our network.

It is intuitive that closeness has the smallest spread (and betweenness right after it) since the closeness for a given vertex is based on the mean shortest distance from it to *every other* vertex. In a network like this, the different between one $s - t$ shortest distance and another is itself not that large since there are many central points connecting things, and certainly the difference between the mean shortest distance of one vertex to all other vertices vs another vertex's mean shortest distance to all vertices is going to be small since it's almost like an average across the whole network with only a small advantage for more connected nodes.

i) Relative ordering:

The only pair of centrality metrics that have the same ordering is {eigenvector, Katz}, which we expected. However, all metrics have Benjamin as the largest, and Gad as smallest then Levi and followed by either Reuben or Yosef as third-smallest. All but "closeness" have either Issachar or Judah as second-largest.

3 Vertex Groups

a) K-Cliques:

The three 3-cliques are:

- A, B, D, I, J, N, R, S, Y, Z
- A, B, D, I, J, L, N, R, S, Y
- A, B, D, G, I, J, N, Y, Z

The one 4-clique is

- A, B, D, G, I, J, L, N, R, S, Y, Z

b) 2-plex:

One such 2-plex is:

- A, B, D, N.

c) K-Cores:

The 2-core is

- A, B, D, I, J, N, R, S, Y, Z

There are no 3-cores or 4-cores.

d) K-Components

One of the 2-components is

- G, I

There are no 3-components

e) Global Clustering Coefficient

The network global clustering coefficient is 0.146341.

f) Local Clustering Coefficient

	A	B	D	G	I	J	L	N	R	S	Y	Z
L. Clust	0	.133	0	0	0	0.333	0	0	1	0.667	0	0

The Watts-Strogatz clustering coefficient is the mean of the local clustering coefficients which is equal to .1778 in this case. This is **not equal** to the global clustering coefficient.

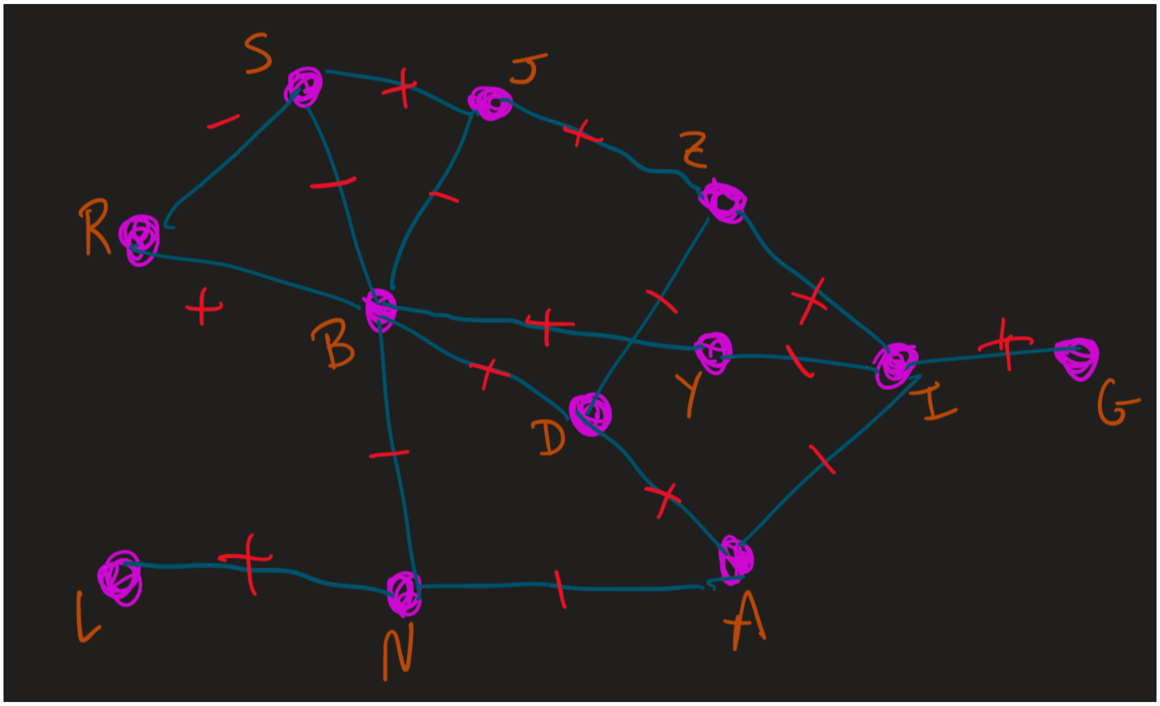
g) Degree and Local Clustering Coefficient

Name	Degree	Local Clust Coeff
Asher	3	0
Benjamin	6	.133
Dan	3	0
Gad	1	0
Issachar	4	0
Judah	3	.333
Levi	1	0
Naphtali	3	0
Reuben	2	1
Simeon	3	.667
Yosef	2	0
Zebulun	3	0

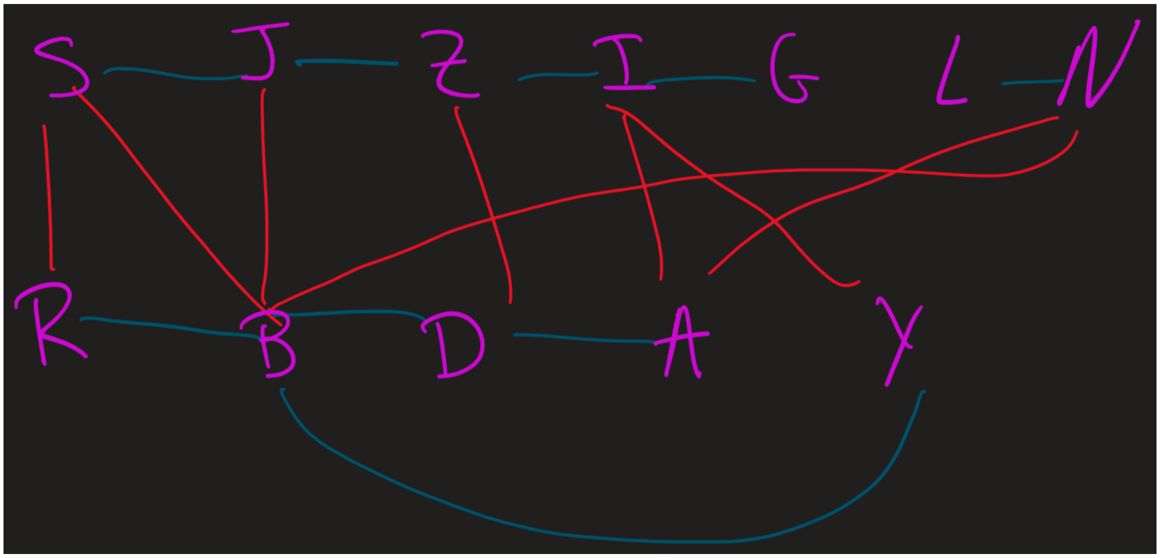
It's interesting that this does not follow the correlation between degree and clustering coefficient as discussed in class. It could be because there are only two such triangles. (BRS and BSJ). Unsurprisingly, those are the only nodes with nonzero clustering coefficients.

4 Signed Networks

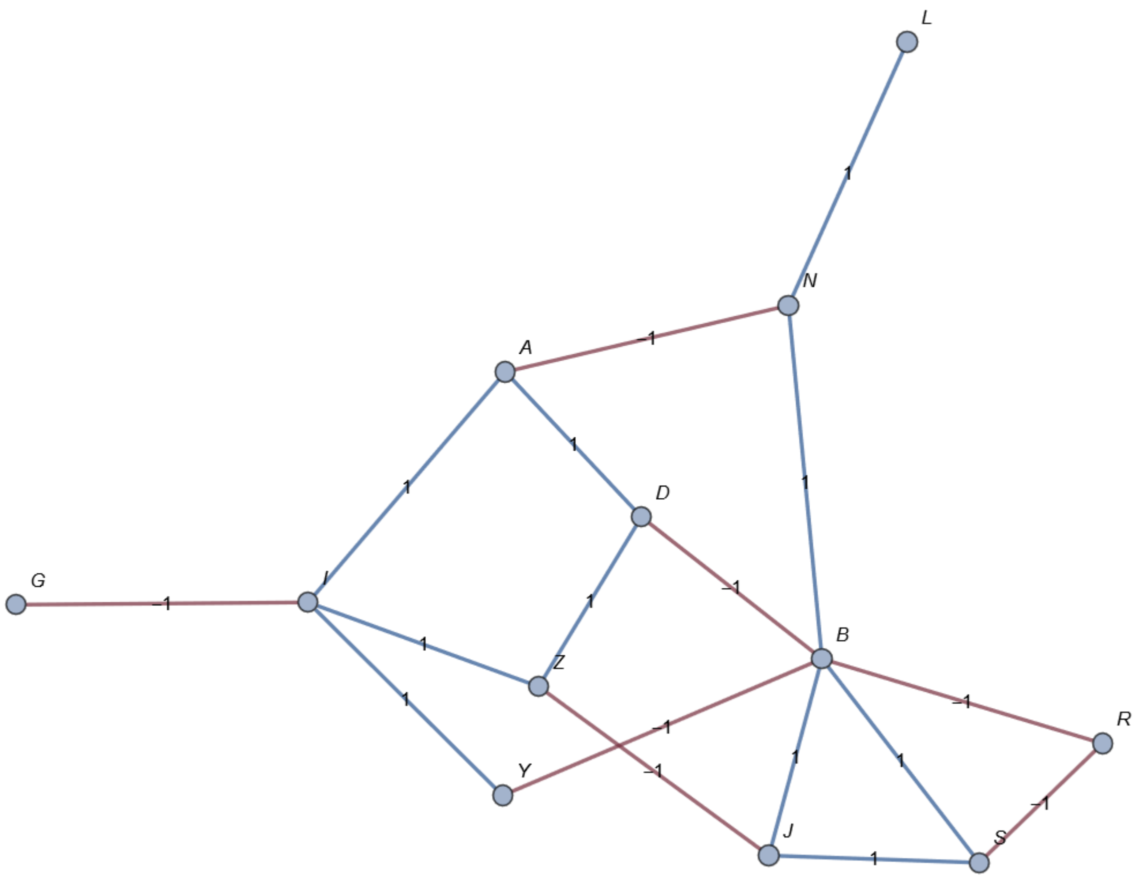
- Yes, there are many such configurations. Here's one:



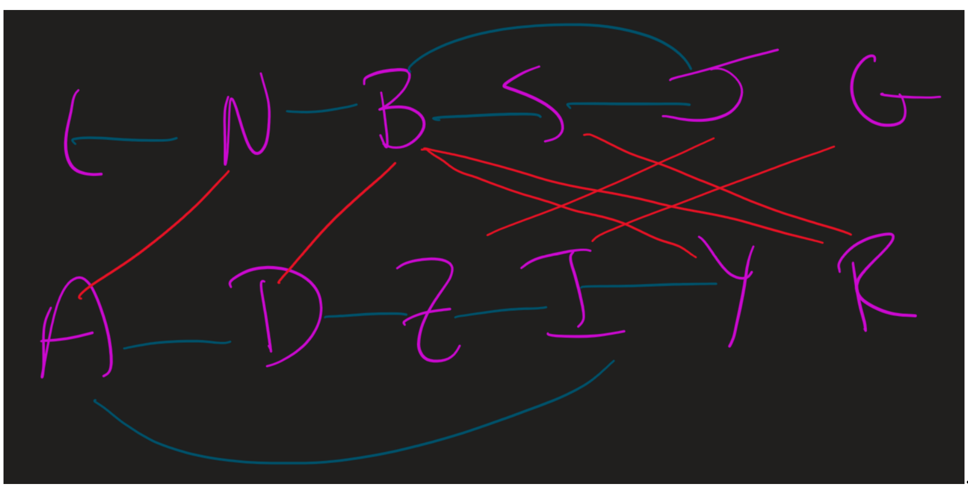
- The corresponding grouping is:



- I found 228 such assignations (not sure if there are doubles there somehow) with 10 (+) bonds and 7 (-) bonds. I picked a particularly aesthetically pleasing one, shown here:



Here is the clusterization:



.

The reason this doesn't violate Harary's Theorem is due to the two "free agent" nodes G and L, which allow us extra flexibility in picking signs since they're not part of any loop.

5 Similarities

a) Cosine Similarities

	A	B	D	G	I	J	L	N	R	S	Y	Z
A	1.	.471	0.	.577	0.	0.	.577	0.	0.	0.	.408	.667
B		1.	0.	0.	.204	.236	.408	0.	.289	.471	0.	.471
D			1.	0.	.577	.667	0.	.667	.408	.333	.408	0.
G				1.	0.	0.	0.	0.	0.	0.	.707	.577
I					1.	.289	0.	.289	0.	0.	0.	0.
J						1.	0.	.333	.816	.333	.408	0.
L							1.	0.	0.	0.	0.	0.
N								1.	.408	.333	.408	0.
R									1.	.408	.5	0.
S										1.	.408	.333
Y											1.	.408
Z												1.

Note that pair repetitions such as LN and NL have been removed, ergo the output table is an upper triangle.

The most similar nodes are J and R with a cosine similarity of .816. My best guess is that it's because they both have a fairly small degree overall and share 2 vertices (part of the same triangle group as in question 3g).

There are many pairs with cosine similarity 0. This is because they share no common vertices.

b) Pearson Correlation Coefficient:

	A	B	D	G	I	J	L	N	R	S	Y	Z
A	1.	.19	-.33	.52	-.41	-.33	.52	-.33	-.26	-.33	.26	.56
B		1.	-.58	-.30	-.35	-.19	.30	-.58	0.	.19	-.45	.19
D			1.	-.17	.41	.56	-.17	.56	.26	.11	.26	-.33
G				1.	-.21	-.17	-.09	-.17	-.13	-.17	.67	.52
I					1.	0.	-.21	0.	-.32	-.41	-.32	-.41
J						1.	-.17	.11	.77	.11	.26	-.33
L							1.	-.17	-.13	-.17	-.13	-.17
N								1.	.26	.11	.26	-.33
R									1.	0.26	0.4	-.26
S										1.	.26	.11
Y											1.	.26
Z												1.

Note that pair repetitions such as LN and NL have been removed, ergo the output table is an upper triangle.

Again the most similar pair is J and R at .77. There is a tie for most dissimilar at -.58 between BD and BN. B is a very well connected node and is actually connected directly to both D and N, but neither of those share any vertices in common with B and each have a two other nodes they're connected to.

6 Assortativity

a) Assortativity

  (a) Leah vs others:
  We get a coefficient of .05556 for the sons of Leah vs the other brothers.

  (b) Six older vs others:
  We get a coefficient of -.0737 for the six older sons vs the other brothers.

b) Assortativity Per Degree:

The Assortativity degree of the network overall is -.37.

7 Directed Networks

a) Reciprocity Degree:

The reciprocity degree of the network is .381

b) Strongly Connected Components:

  (a) A,D
  (b) G
  (c) Y
  (d) L
  (e) N
  (f) B, I, J, Z
  (g) S
  (h) R

c) Ins and Outs

  (a) out: {A, D}, in: {A, D, I, N, B, Z, J, R, S}
  (b) out: {G}, in: {G, I, Z, J, B, S, R}
  (c) out: {Y}, in: {Y, B, I, J, R, S, Z}
  (d) out: {L}, in: {L, N, B, J, R, S, Z, I}
  (e) out: {N, A, L, D}, in: {N, B, J, R, S, Z, I}

(f) out: {B, N, J, D, Y, A, L, Z, I, G}, in: {B, J, R, S, Z, I}

(g) out: {S, B, J, N, D, Y, Z, A, L, I, G}, in: {S, R}

(h) out: {R, B, S, N, J, D, Y, A, L, Z, I, G}, in: {R}

d) Shortest Distance:

Here I pick node B because it seems like it will be interesting since it's connected to so many other nodes.

	Undirected	To B	From B
A	2	∞	2
B	0	0	0
D	1	∞	1
G	3	∞	4
I	2	3	3
J	1	1	1
L	2	∞	2
N	1	∞	1
R	1	1	∞
S	1	1	∞
Y	1	∞	1
Z	2	2	2

Where a path length of ∞ means that it is impossible to get from S to T, in our case because of the direction of the edges and such. We have undirected which is the ST path B-i for every i row (BA, BB, BD etc). Then we have "To B" which is the paths *row*-B (AB, BB, DB etc), and "From B" which is B-*row* (BA, BB, BD etc.).

For the most part, if a path is possible in the directed network it's roughly the same length as in the undirected case. For example, for B to I instead of going directly through Y (B-Y-I), we need to go through J and Z (B-J-Z-I) and exactly the same in reverse since Z and J both have in and out edges, whereas Y is a pure sink and therefore we can't travel through it.

There are 8 paths which are not reachable, mostly To B since there's only one in-edge to B, from J.

e) Bonus Question:

	Undirected	To B	From B	Weighted
A	2	∞	2	2
B	0	0	0	0
D	1	∞	1	1
G	3	∞	4	3
I	2	3	3	2
J	1	1	1	2
L	2	∞	2	2
N	1	∞	1	1
R	1	1	∞	1
S	1	1	∞	1
Y	1	∞	1	1
Z	2	2	2	2

It seems like I picked a fairly unfortunate node since the only difference between the undirected + unweighted and this new weighted graph is from B to J where the shortest route is directly through the double edge with a cost of 2 (whereas in every other case – undirected, from B, and to B – the distance is 1 since it's accessible both ways).