## Sociology 375 Exam 1 Fall $2012 \quad$ Prof Montgomery

Answer all questions. 250 points possible. Explanations can be brief.

1) [70 points] Consider the relation A ('sends information to') on a set of four actors $S=$ $\{1,2,3,4\}$. In particular, suppose that 1 sends information to 2,2 sends information to 1 , 2 sends information to 3 , 3 sends information to 4 , and 4 sends information to 3 .
a) Show how the relation A could be represented as
i) a set
ii) a directed graph
iii) an adjacency matrix
b) Determine (by inspection or computation)
i) the number of 3-paths between each ordered pair of actors
ii) the reachability matrix
iii) the distance matrix
c) In this example, is the reachability relation an equivalence relation? If so, find the equivalence classes generated by this relation. If not, explain why.
d) Suppose we partition the set of actors into those actors who can be reached by all others and those actors who cannot. Which actors are in the first subset? Which are in the second subset? Use a matrix text to determine whether this partition is a regular equivalence. If the matrix test fails, briefly explain the "problem" with this partition as a regular equivalence.
2) [40 points] An Olympic tournament included teams from America (A), Britain (B), Canada (C) and Denmark (D). In the first round, A beat B, and C beat D. There was then a "gold medal" game where A beat C, and a "bronze medal" game where B beat D.
a) Give the adjacency matrix for the 'beat' relation. Use a matrix test to determine whether this relation is transitive. Is the beat relation a strict partial order? Explain.
b) State the definition for "structural equivalence" between two actors i and j. Given the beat relation, partition the teams into structural equivalence classes.
c) Based your preceding answers, what problems arise in awarding the gold medal (first place) to A, the silver medal (second place) to C, the bronze medal (third place) to B, and no medal (last place) to D ? Would the ranking necessarily become clearer if the tournament also included a game between A and D , and a game between B and C ? Briefly explain.
3) [70 pts] Consider a graph with 100 nodes indexed $\mathrm{i} \in\{1,2,3, \ldots, 100\}$.
a) Suppose the graph takes the form of a ring where each node $i$ is connected to nodes $i-1$ and $\mathrm{i}+1$. (To close the ring, assume that node 100 is connected to nodes 99 and 1 , and that node 1 is connected to nodes 100 and 2.) What is the clustering coefficient for this graph? What is the diameter of this graph? What is the average distance over all pairs of nodes? [HINT: Given the symmetry of the ring structure, all nodes have the same clustering coefficient. I will accept approximate answers for diameter and average distance, but you need to explain your reasoning to receive credit.]
b) Consider a similar ring structure where each node $i$ is connected to nodes $i-3, i-2, i-1$, $\mathrm{i}+1$, $\mathrm{i}+2$, and $\mathrm{i}+3$. (To close the ring, assume node 98 is connected to nodes $95,96,97$, 99,100 , and 1 . Node 99 is connected to nodes $96,97,98,100,1$, and 2 . And so on.) What is the clustering coefficient for this graph? What is the diameter? What is the average distance over all pairs of nodes? [HINT: Same as hint for part (a).]
c) Suppose we modify the ring structure in part (a) by adding a "shortcut" (i.e., a direct connection) between nodes 1 and 51. Does this shortcut reduce the diameter of the graph? If so, what is the new diameter? If not, explain why. Does the shortcut reduce average distance by more than half, about half, or less than half? Explain why.
d) Suppose we start with the ring structure in part (b), and then begin to replace randomly chosen edges in the ring structure with "shortcuts" between randomly chosen pairs of nodes. How would the clustering coefficient and average distance change as we continue to replace the original edges in the ring structure with shortcuts? Explain how this thought experiment addresses the "small world" phenomenon.
4) [70 points] Consider the graph below. To answer the following questions, it may be helpful to use the adjacency matrix, distance matrix and connectivity matrix given at the end of the problem.

a) Is $\{1,3,4,5\}$ a (strong) clique? Explain why or why not. Then find all (strong) cliques with at least 3 members.

4b) Is $\{1,2,3,4,5\}$ a 2-clique? Explain why or why not. Then find all the 2-cliques. [HINT: There are 2 of them.]
c) Are both of the 2-cliques also 2-clans? Explain why or why not.
d) Find any 2-clubs that are not 2-cliques. [HINT: There are 2 of them.]
e) In general, are $\mathrm{n}+1$-cliques more cohesive or less cohesive than n -cliques? Are $\mathrm{k}+1$ components more cohesive or less cohesive than k-components?
f) What is the connectivity level $k$ of the entire graph shown above? What are the "global" implications of this connectivity level? Illustrate by finding a k-cutset for the graph. What are the "local" implications of this connectivity level? Illustrate using the pair of nodes $\{2,6\}$.

```
>> A
A =
    0
    1
    1
    1
    1
    0
>> distance(A)
ans =
    0
    1
    1
    1
    1
    2
>> con = [ ]; for i = 1:7; for j = 1:7; con(i,j) = connectivity(A,i,j); end; end; con
con =
    Inf Inf Inf Inf Inf 2 2
    Inf Inf 2 2 Inf 2 2
    Inf 2 Inf Inf 3 Inf 2
    Inf 2 Inf Inf Inf 2 2
    Inf Inf 3 Inf Inf 2 Inf
        2 2 Inf 2 2 Inf Inf
        2 2 2 2 Inf Inf Inf
```

1a) $[15 \mathrm{pts}] \quad \mathrm{A}=\{(1,2),(2,1),(2,3),(3,4),(4,3)\}$


$$
A=\left[\begin{array}{llll}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{array}\right]
$$

b) $[25 \mathrm{pts}]$

$$
\begin{aligned}
& \mathrm{A}^{3}=\left[\begin{array}{llll}
0 & 1 & 0 & 1 \\
1 & 0 & 2 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{array}\right] \\
& \text { reachability }=\mathrm{R}=\left(\mathrm{I}+\mathrm{A}+\mathrm{A}^{2}+\mathrm{A}^{3}\right) \#=\left[\begin{array}{llll}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right] \\
& \text { distance }=\left[\begin{array}{llll}
0 & 1 & 2 & 3 \\
1 & 0 & 1 & 2 \\
\infty & \infty & 0 & 1 \\
\infty & \infty & 1 & 0
\end{array}\right]
\end{aligned}
$$

c) [10 pts] In this example, the reachability relation is reflexive and transitive but not symmetric. (Note that 1 and 2 can reach 3 and 4 , but 3 and 4 cannot reach 1 or 2.) Thus, reachability is not an equivalence relation.
d) [20 pts] Given the $R^{T}$ matrix we see that actors $\{3,4\}$ can be reached by all while $\{1,2\}$ cannot. This partition can be represented as an adjacency matrix E (see below), and is a regular equivalence of the relation A if AE \# = EA\#. Computing these matrices, we find that

$$
\mathrm{AE}=\left[\begin{array}{llll}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{array}\right]\left[\begin{array}{llll}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right]=\left[\begin{array}{llll}
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right]
$$

$$
\mathrm{EA}=\left[\begin{array}{llll}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right]\left[\begin{array}{llll}
0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{array}\right]=\left[\begin{array}{llll}
1 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right]
$$

Because $\mathrm{AE} \# \neq \mathrm{EA} \#$, this partition is not a regular equivalence. Regular equivalence is violated because 1 does not know someone like 3, but someone like 1 does know 3 (i.e., AE\# $(1,3)=0$ but EA\#(1,3) = 1), and also because 2 knows someone like 4 but someone like 2 does not know 4 (i.e., $\operatorname{AE\# }(2,4)=1$ but $\operatorname{EA\# }(2,4)=0)$.

2a) [20 pts] Let $R$ denote the adjacency matrix for the 'beat' relation.

$$
\mathrm{R}=\left[\begin{array}{cccc}
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{array}\right] \quad \mathrm{R}^{2}=\left[\begin{array}{cccc}
0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}\right] \quad \mathrm{R}-\mathrm{R}^{2} \#=\left[\begin{array}{cccc}
0 & 1 & 1 & -1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{array}\right]
$$

Transitivity requires $\left(R-R^{2} \#\right) \geq 0$ (i.e., every element of the test matrix must be nonnegative). Thus, the beat relation is not transitive. Strict partial orders must be irreflexive, asymmetric, and transitive. Thus, the beat relation is not a strict partial order.
b) [10 pts] Actors i and j are structurally equivalent when they send ties to and receive ties from the same others. Equivalently, given the adjacency matrix for the relation R, rows i and j and columns i and j must be the same. In the present example, the structural equivalence classes are $\{A\},\{B, C\}$, and $\{D\}$.
c) [10 pts] Because B and C are structurally equivalent (each team lost to A and beat D), there is no good reason to rank one ahead of the other. Further, while the paths $\mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{D}$ and $\mathrm{A} \rightarrow \mathrm{C} \rightarrow \mathrm{D}$ would suggest that A should be ranked above D , there was no game between $A$ and $D$ to verify transitivity. A game between $B$ and $C$ would allow comparison between those two teams, and a transitive ordering of $\mathrm{A}, \mathrm{B}$, and C . If there was also a game between A and D, and if A was the winner, there would be a linear order (i.e., a complete strict partial order) over all teams. On the other hand, if D beat A, this would create a cycle in the graph, and there would be no clear ordering of teams.

3a) [15 pts] Each node i has 2 neighbors (so there is one possible tie between i's neighbors), and these neighbors don't know each other (so there are no actual ties). Thus, each node has a clustering coefficient of $0 / 1=0$. The clustering coefficient for the graph is the mean of all the individual clustering coefficients, and so is also 0 . The diameter (largest distance between pairs) in the graph is 50 . From any node i, the highest distance is 50 (to the "other side" of the ring) and the lowest distance is 0 (to node i) and the average distance across all other nodes is 25 . Thus, average distance for the graph is 25 .

3b) [20 pts] Each node i has 6 neighbors, who potentially have $6 * 5 / 2=15$ ties among themselves. As shown below, 9 of these ties are actually present.


Thus, the clustering coefficient for each node is 9/15. Again, because the clustering coefficient is the same for every node, the clustering coefficient for the graph is also 9/15. The shortest path to the "other side" of the ring (e.g., from node 1 to node 51) is now 17. (Consider the path $1 \rightarrow 4 \rightarrow 7 \rightarrow \ldots \rightarrow 49 \rightarrow 51$. Note that path length is $50 / 3$ rounded upward.) Average distance from any node to others is approximately $17 / 2=8.5$.
c) [20 pts] No, the diameter of the graph is still 50. The distance between nodes 26 and 76 is still 50. The "shortcut" reduces average distance between node 1 (or 51) and other nodes by half. Average distance thus falls by $50 \%$ from nodes 1 or 51 , doesn't fall at all for nodes 26 or 76 , and falls between 0 and $50 \%$ for other nodes. Overall, average path length falls by about $25 \%$ (which is less than half).
d) [15 pts] This thought experiment is similar to the numerical experiment you performed on Problem Set 5. Let t denote the number of edges removed from the ring structure and added randomly. As t rises, the average distance will initially fall rapidly and then fall more slowly (i.e., non-linearly), while the clustering coefficient will fall slowly and steadily (i.e., linearly). Thus, for relatively low values of $t$, we obtain a "small world" graph in which average distance is low (i.e., there are short paths between most pairs) while the clustering coefficient is high (i.e., the graph remains highly "structured").
[[[NOTE: You didn’t need Matlab to complete problem 3. But if you want to experiment with this ring structure yourself, you can create create the adjacency matrix using the code below. The variable $s$ is the number of neighbors on each side of each node.
$\gg \mathrm{s}=1 ; \mathrm{A}=\operatorname{zeros}(100)$; for $\mathrm{i}=1: 100 ; \mathrm{A}(\mathrm{i}, \bmod (\mathrm{i}: \mathrm{i}+\mathrm{s}-1,100)+1)=1$; end; $\mathrm{A}=\mathrm{A} \mid \mathrm{A}^{\prime}$
4a) [14 pts] No, not a strong clique because the subgraph is not complete: there is no edge between 3 and 5 . The cliques are $\{1,2,5\},\{1,3,4\}$, and $\{1,4,5\}$.
b) [15 pts] No, not a 2-clique because this set of nodes is not maximal (with respect to the property "all pairs have distance $\leq 2$ "). In particular, you could add node 7 to the set to obtain the 2 -clique $\{1,2,3,4,5,7\}$. From inspection of the distance matrix, you cannot include both 2 and 6 in the same 2 -clique. The other 2 -clique is $\{1,3,4,5,6,7\}$.
c) [5 pts] No, $\{1,2,3,4,5,7\}$ is not a 2-clan because the distance between 3 and 7 is greater than 2 within the subgraph.
d) [10 pts] Removing 3 from the 2-clique in part (c), we obtain the 2 -club $\{1,2,4,5,7\}$. Removing 7 from this 2-clique, we obtain the 2-club $\{1,2,3,4,5\}$.
e) $[6 \mathrm{pts}] \mathrm{n}+1$-cliques are less cohesive than n -cliques. (Recall that 1-cliques are strong cliques.) $\mathrm{k}+1$-components are more cohesive than k -component. (Recall that k is the size of the smallest cutset, making $\mathrm{k}+1$-components more robust to loss of members.)
f) [20 pts] The graph is a 2-component (because 2 is the smallest number in the connectivity matrix). Globally, we would need remove at least 2 nodes to break the graph into components. The 2-cutsets are $\{1,5\},\{3,5\},\{3,7\},\{5,6\}$. Locally, each pair of nodes cannot be separated by cutsets with fewer than 2 members. Further, there must be at least 2 node-independent paths between each pair. The pair $\{2,6\}$ could be separated by any of the first three cutsets listed above (but not by any 1-cutset). Two node-independent paths are $(2,1,3,6)$ and $(2,5,7,6)$.

## Sociology 375 Exam 2 Fall 2012 Prof Montgomery

Answer all questions. 260 points possible.

1) [75 points] The following data matrix indicates whether actors $i \in\{1,2,3,4,5\}$ participated in events $j \in\{A, B, C\}$. $D(i, j)=1$ if actor $i$ attended event $j$, and $D(i, j)=0$ otherwise.

$$
\mathrm{D}=\left[\begin{array}{lll}
1 & 0 & 1 \\
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1 \\
0 & 0 & 1
\end{array}\right]
$$

a) We may say that actor i "contains" actor k if $\mathrm{D}(\mathrm{i}, \mathrm{j}) \geq \mathrm{D}(\mathrm{k}, \mathrm{j})$ for all events j . Give this containment relation in matrix form, and then draw the Hasse diagram. [HINT: If you use matrix multiplication to derive the Hasse matrix, you should remove the 1 s from the main diagonal of the containment matrix to create a strict partial order. But it may be much faster to find both the containment matrix and Hasse diagram by inspection.]
b) Construct the (exact) Galois lattice for the D matrix. Use reduced labeling (for both actors and events).
c) Using the HICLAS method to obtain a rank-2 approximation of the D matrix, we obtain the S matrix (of "row bundles") and the P matrix (of "column bundles") below.

$$
\mathrm{S}=\left[\begin{array}{ll}
1 & 0 \\
1 & 1 \\
1 & 1 \\
0 & 1 \\
0 & 0
\end{array}\right] \quad \mathrm{P}=\left[\begin{array}{ll}
1 & 0 \\
0 & 1 \\
0 & 0
\end{array}\right]
$$

What is the equation for the estimated matrix M (as a function of the S and P matrices)? Compute this estimated matrix M for the S and P matrices above. Comparing the estimated matrix M to the data matrix D , how many discrepancies are generated by the HICLAS estimate?
d) Construct the (approximate) Galois lattice using the S and P matrices.
e) Using the HICLAS method, could other rank-2 approximations of D generate more discrepancies than found in part (c)? Could other rank-2 approximations of D generate fewer discrepancies? How many discrepancies would be generated by the best rank-3 approximation? Briefly explain.
2) [45 points] Given the D matrix from question 1 , we can conduct a (very simple) factor analysis. The following matlab computations are necessary.

```
>> D
D =
    1 0}
    1 1 0
    1 1 1
    0}11
    0}00
>> X = corrcoef(D)
X =
    1.0000 0.1667 -0.4082
    0.1667 1.0000 -0.4082
    -0.4082 -0.4082 1.0000
>> [Y,Z] = eig(X)
Y =
    0.4629 0.7071 -0.5345
    0.4629 -0.7071 -0.5345
    0.7559 0.0000 0.6547
Z =
    0.5000 0 0
            0}0.8333 
            0}001.666
```

a) What are the Y and Z matrices?
b) Following the usual procedure in factor analysis, draw a 2-dimensional scatterplot. [HINT: Given that you're doing this by hand, your scatterplot doesn't need to be perfect, but you should label the points and axes appropriately, and indicate numerical coordinates. The labels for the points should be taken from question 1.]
c) In general, what is the purpose of factor analysis? How would the present analysis differ if we began with the correlation-coefficient matrix for the transpose of the D matrix (i.e., if we set $\mathrm{X}=\operatorname{corrcoef}\left(\mathrm{D}^{\prime}\right)$ )?
d) What is the "trace" of a matrix? How is the trace of a (square, symmetric) matrix related to its eigenvalues?
e) State the standard formula in the factor-analysis literature for the "proportion of variance" explain by each factor. For the example above, what proportion of variance is explained by the first factor? By the second factor? By the first two factors combined?
3) [55 points] Each of the matrices below characterize positive and negative relations on a set of individuals $\{1,2,3,4\}$. In each case, $M(i, j)=1$ if $i$ and $j$ are friends, $M(i, j)=-1$ if $i$ and $j$ are enemies, and $M(i, j)=0$ if $i$ and $j$ are neither. Note that these relations are symmetric so that $M(i, j)=M(j, i)$ for all pairs (i,j). By convention, $M(i, i)=0$ for all $i$.

For each case, determine whether the signed graph characterized by the matrix M is balanced. If it is balanced, you should report the relevant partition of actors. Otherwise, you should explain why the graph is not balanced, and then determine whether it is clusterable. If it is clusterable, you should report the relevant partition of actors. Otherwise, you should explain why the graph is not clusterable.

$$
\begin{array}{ll}
\text { (a) } M=\left[\begin{array}{cccc}
0 & -1 & -1 & -1 \\
-1 & 0 & -1 & 1 \\
-1 & -1 & 0 & -1 \\
-1 & 1 & -1 & 0
\end{array}\right] & \text { (b) } M=\left[\begin{array}{cccc}
0 & 1 & -1 & 1 \\
1 & 0 & 0 & -1 \\
-1 & 0 & 0 & 1 \\
1 & -1 & 1 & 0
\end{array}\right] \\
\begin{array}{ll}
\text { (c) } M=\left[\begin{array}{cccc}
0 & -1 & 1 & 0 \\
-1 & 0 & -1 & -1 \\
1 & -1 & 0 & 1 \\
0 & -1 & 1 & 0
\end{array}\right] & \text { (d) } M=\left[\begin{array}{cccc}
0 & 1 & -1 & -1 \\
1 & 0 & 1 & 1 \\
-1 & 1 & 0 & 1 \\
-1 & 1 & 1 & 0
\end{array}\right]
\end{array}>.
\end{array}
$$

4) [30 points] Consider a kinship system in a society with clans $\{1,2,3\}$ characterized by the matrices

$$
\mathrm{W}=\left[\begin{array}{lll}
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{array}\right] \quad \mathrm{C}=\left[\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right]
$$

where $\mathrm{W}(\mathrm{i}, \mathrm{j})=1$ indicates that a man in clan i must marry a woman in clan j , and $\mathrm{C}(\mathrm{i}, \mathrm{j})=1$ indicates that a man in clan i has children in clan j .
a) Assuming that you are never allowed to marry within your own clan, there is no kinship system in which a man is allowed to marry his father's brother's daughter nor his mother's sister's daughter. Explain why, using the relevant matrix equations.
b) If every man is allowed to marry his mother's brother's daughter, what matrix equation must be satisfied? Is this condition satisfied given the W and C matrices above?
c) If every man is allowed to marry his father's sister's daughter, what matrix equation must be satisfied? Is this condition satisfied given the W and C matrices above?
5) [55 points] Consider the following relation on the set of actors $\{1,2,3,4,5\}$ :

$$
R=\{(1,2),(1,3),(1,5),(3,1),(4,5),(5,3),(5,4)\}
$$

a) Draw the graph of this relation. What is the total number of dyads in this graph? What is the total number of triads?
b) Using your graph, compute the dyad census and the triad census. [HINT: Following Davis-Leinhardt-Holland, the three types of dyads are M and A and N. While the dyad census can be computed using matrix methods, it will be much faster to compute this census by inspection. The 16 possible types of triads are given on the attached sheet (Holland and Leinhardt, 1971, Figure 3, p 118). While the triad census can be computed using matrix methods, it will be much faster to list every triad from the graph in part (a) and then determine each triad's type.]
c) Following Davis-Leinhardt-Holland, we may use the dyad census to determine the expected number of each type of triads. Using the dyad census from part (b) and assuming sampling with replacement, what is the expected number of triads of type 003 ? of type 012 ? of type 300 ? Given the expected number of type 300 (and looking again at your dyad census), why would sampling without replacement seem more appropriate for deriving the expected numbers of triads?

Soc 375 Exam 2 Fall 2012 Solutions
1a) [15 pts] The (weak) containment matrix and Hasse diagram are
$\gg \mathrm{C}=\sim\left(\sim \mathrm{D}^{*} \mathrm{D}^{\prime}\right)$
$\mathrm{C}=$
$\begin{array}{lllll}1 & 0 & 0 & 0 & 1\end{array}$
$\begin{array}{lllll}0 & 1 & 0 & 0 & 0\end{array}$
$\begin{array}{lllll}1 & 1 & 1 & 1 & 1\end{array}$
$\begin{array}{lllll}0 & 0 & 0 & 1 & 1\end{array}$
$0 \begin{array}{lllll}0 & 0 & 0 & 0 & 1\end{array}$


5
b) [20 pts] Closing the rows of the D matrix under intersection, we obtain the expanded

D matrix:
$\mathrm{D}=$
$1 \quad 0 \quad 1$
110
$1 \quad 1 \quad 1$
$0 \quad 1 \quad 1$
$0 \quad 0 \quad 1$
$1 \quad 0 \quad 0 \quad$ (intersection of rows 1 and 2)
$\begin{array}{llll}0 & 1 & 0 & \text { (intersection of rows } 2 \text { and 4) }\end{array}$
$0 \quad 0 \quad 0 \quad$ (intersection of rows 2 and 5)
The containment relation on rows of the expanded D matrix determines the edges of the lattice. Using reduced labeling,

(To check this answer, you could derive the original D matrix from the lattice.)

1c) [20 pts] The estimated matrix is given by the equation

$$
M=\sim\left(\sim S * P^{\prime}\right)
$$

where $\sim$ denotes complement and $*$ denotes Boolean multiplication. (In matlab, the outer $\sim$ converts the expression in parenthesis into a binary matrix, so you obtain the correct answer using standard matrix multiplication).

```
>> M = ~(~S*P')
M =
    1 0
    1 1 1
    1 1 1
    0}11
    0}00
```

Comparing the estimated matrix M to the data matrix D , there is only 1 discrepancy: $\mathrm{M}(2,3)=1 \neq \mathrm{D}(2,3)=0$.
d) [10 pts] The lattice for the rank-2 approximation is

e) [10 pts] We could certainly find worse rank-2 approximations (generating more discrepancies), but cannot reduce the number of discrepancies to 0 unless we move to a rank-3 approximation. (Recall that the exact Galois lattice in part b was a full rank-3 lattice.) Because the original D matrix has 3 columns, we can find an exact rank-3 solution (by setting $\mathrm{S}=\mathrm{D}$ and $\mathrm{P}=\mathrm{I}$ ).

2a) [5 pts] Each column of $Y$ is an eigenvector of $X$. The associated eigenvalue is placed on the main diagonal of Z .

2b) $[10 \mathrm{pts}]$


2c) [10 pts] Essentially, factor analysis provides a low-dimensional (usually 2dimensional) spatial representation of the information in the correlation matrix. Variables that are highly correlated will be placed near each other on the scatterplot. The present analysis uses the correlations between column variables (the events A,B,C). If we used the transpose of the D matrix, we would have considered the correlations between the actors (and the points in the scatterplot would have been labeled 1,2,3,4,5).
d) [5 pts] The trace is the sum of the diagonal elements. For a square, symmetric matrix, the trace is equal to the sum of the eigenvectors. [For the present example, note that $\operatorname{trace}(\mathrm{X})=\operatorname{trace}(\mathrm{Z})$.]
e) [15 pts] The proportion of variance explained by factor i is equal to

$$
\lambda_{\mathrm{i}} / \operatorname{trace}(\mathrm{Z})=\lambda_{\mathrm{i}} / \operatorname{trace}(\mathrm{X})=\lambda_{\mathrm{i}} / \mathrm{n}
$$

where $\lambda_{\mathrm{i}}$ is the eigenvalue associated with factor $\mathrm{i}, \mathrm{Z}$ is the diagonal eigenvalue matrix, and X is the $\mathrm{n} \times \mathrm{n}$ correlation matrix. For the current problem, the first factor explains $(1.666) / 3=55.55 \%$ of the variance, while the second factor explains $(.833) / 3=27.77 \%$ of the variance. Thus, the first two factors combined explain (1.666+.833)/3=(2.5)/3= 83.33\% of the variance.

3a) [15 pts] Not balanced. Negative cycle (1,3,4,1). Clusterable into $\{\{1\},\{3\},\{2,4\}\}$.
b) [15 pts] Not balanced. Not clusterable. Cycle with one negative edge (1,3,4,1).
c) $[10 \mathrm{pts}]$ Balanced. Partition is $\{\{1,3,4\},\{2\}\}$.
d) [15 pts] Not balanced. Not clusterable. Cycle with one negative edge (1,2,4,1).

4a) [10 pts] A man may marry his father's brother's daughter if $\mathrm{C}^{-1} \mathrm{C}=\mathrm{W}$, and may marry his mother's sister's daughter if $\mathrm{C}^{-1} \mathrm{WW}^{-1} \mathrm{C}=\mathrm{W}$. But since the left-hand-side of each of these equations reduces to I , and since $\mathrm{W} \neq \mathrm{I}$, these types of marriages are never allowed.
b) $[10 \mathrm{pts}]$ The condition is $\mathrm{C}^{-1} \mathrm{WC}=\mathrm{W}$, often written as $\mathrm{WC}=\mathrm{CW}$. In this example, because $\mathrm{C}=\mathrm{I}$, it is obvious that this equation holds.
c) [10 pts] The condition is $\mathrm{C}^{-1} \mathrm{~W}^{-1} \mathrm{C}=\mathrm{W}$, often written as $\mathrm{W}^{-1} \mathrm{C}=\mathrm{CW}$. In this example, because $\mathrm{C}=\mathrm{I}$, this condition reduces to $\mathrm{W}^{-1}=\mathrm{W}$, which is not true.

5a) [15 pts]


Dyads are unordered pairs. Given 5 nodes, there are (5*4)/2 $=10$ dyads Triads are unordered triples. Given 5 nodes, there are $(5 * 4 * 3) /(3 * 2)=10$ triads.
b) [20 pts] By inspection of the graph, we obtain $M=2, A=3, N=5$. (You could also obtain this result by listing every dyad - \{1,2\}, $\{1,3\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\}$, $\{3,4\},\{3,5\},\{4,5\}$ - and determining its type. To conduct the triad census, we can list the 10 triads and then determine the type for each triad.

| $\{1,2,3\}$ | 111 U |
| :--- | :--- |
| $\{1,2,4\}$ | 012 |
| $\{1,2,5\}$ | 021 D |
| $\{1,3,4\}$ | 102 |
| $\{1,3,5\}$ | 120 C |
| $\{1,4,5\}$ | 111 D |
| $\{2,3,4\}$ | 003 |
| $\{2,3,5\}$ | 012 |
| $\{2,4,5\}$ | 102 |
| $\{3,4,5\}$ | 111 U |

Arranging the number of each type of triad in the conventional order (which you didn't need to remember), the triad census is usually written as a vector

T = [T003 T012 T102 T021D T021U T021C T111D T111U T030T T030C T201 T120D T120U T120C T210 T300]

Here, $T=(1,2,2,1,0,0,1,2,0,0,0,0,0,1,0,0)$.
c) [20 pts] From the dyad census, the probabilities of each type are $\mathrm{m}=1 / 5, \mathrm{a}=3 / 10, \mathrm{n}=$ $1 / 2$. Thus, the expected number of type 003 triads would be $10 * n * n * n=1.25$, the expected number of type 012 triads would be $10 * 3 * a * n * n=2.25$, the expected number of type 300 triads would be $10 * m * m * m=.08$. However, recognizing there were only 2 mutual edges in the dyad census, it would obviously be impossible to construct a type 300 triad. This constraint is ignored when sampling with replacement, but would be addressed by sampling without replacement, which implies that the probability of a type 300 triad would be $(2 / 10)(1 / 9)(0 / 8)=0$.

