THE UNIVERSITY *of* ADELAIDE

Examination in School of Mathematical Sciences
Semester 2, 2017

105637	APP MTH 4052	Applied Topic F: Complex Network ...
105661	APP MTH 7088	Applied Topic F: Complex Network ...
108646	STATS 4008	Stats Topic D: Complex Network ...
004013	STATS 7008	Stats Topic D: Complex Network ...

Official Reading Time:	10 mins
Writing Time:	180 mins
Total Duration:	190 mins

NUMBER OF QUESTIONS: 6 TOTAL MARKS: 50

Instructions

- Attempt all questions.
- Begin each answer on a new page.
- Examination materials must not be removed from the examination room.

Materials

- 1 Blue book is provided.
- Calculators without remote communications facilities are permitted.
- Students are permitted to bring two, double-sided pages of handwritten notes.
- English and foreign-language dictionaries may be used.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO.

1. (a) (i) The adjacency matrix is

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

[1 mark]

(ii) The node degrees are the row or column sums (for an undirected graphs) and so are

node	1	2	3	4
degree	3	2	2	3

[1 mark]

(b) (i) The handshake lemma states that the total number of edges in an undirected graph is half the sum of the degrees of the nodes, or

$$2|E| = \sum_{i \in N} k_i,$$

for the graph $G = (N, E)$ with node degrees k_i for all $i \in N$. [1 mark]

(ii) **Proof:** The degree of a node (in an undirected graph) is the number of links which connect to it. Each link connects two nodes, so the sum over all the degrees includes each link twice, and is therefore twice the number of links in the graph. [2 marks]

(iii) In this graph, there are 5 edges, and the sum of the node degrees is $10 = 2 \times 5$. [1 mark]

(c) (i) Local clustering coefficients and the average are shown in the following table

node	k_i	$k_i(k_i-1)/2$	$\lvert\{(j,k) \in E \mid j,k \in N_i\}\rvert$	c_i
1	3	3	2	2/3
2	2	1	1	1
3	2	1	1	1
4	3	3	2	2/3
average				10/12

[2 marks]

(ii) The number of triangles in the graph is 2, and the number of triples is 8. Thus the global clustering coefficient is

$$C = \frac{3 \times 2}{8} = \frac{3}{4}.$$

[1 mark]

Obviously the average of the local coefficients, and the global coefficient are different, but both indicate a high degree of clusterings.

[1 mark]

(d) (i) Shortest paths (N.B. as the links are undirected, we only have to calculate paths in one direction). You can use which-ever algorithm you prefer to calculate these.

OD pair	distance	path
(1,2)	2	1-2
(1,3)	4	1-3
(1,4)	5	1-2-4
(2,3)	5	2-4-3
(2,4)	3	2-4
(3,4)	2	3-4
average	3.5	

[3 marks]

(ii) To calculate the diameter, we can first compute the node eccentricity, which is the longest (shortest-path) distance from that node to any other. The diameter is the maximum of these:

node	distances	eccentricity
1	2,4,5	5
2	2,5,3	5
3	4,5,2	5
4	5,3,2	5
diameter		5

We could also note that the diameter is the longest (shortest) path in the network, and hence it is once again 5.

[1 marks]

2. (a) The graph products are shown in Figure 1.

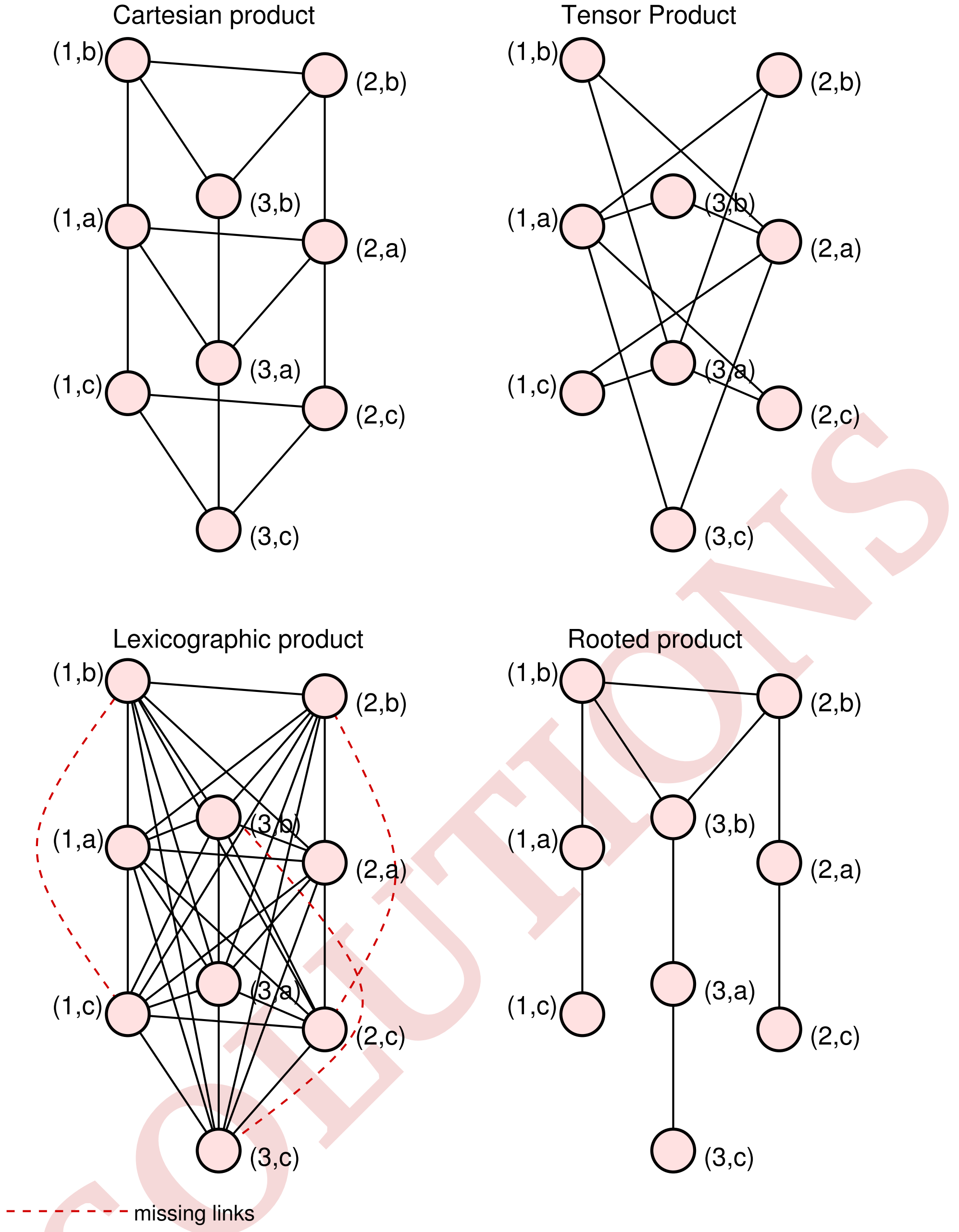


Figure 1: Graph operations results. Note in the lexicographic product case, it is easaier to record the small number of links that aren't present, than the links that are. N.B. You didn't need to do the lexicographic product.

[6 marks]

(b) The Cartesian product commutes in the sense that $G\Box H \simeq H\Box G$, *i.e.*, that the two resulting graphs are isomorphic.

[1 mark]

3. (a) Denote the set of edges from nodes of Type i to those of Type j as E_{ij} (noting that as the network is undirected $E_{ij} = E_{ji}$). Then the average number of each will be given by the number of possible connections times the probabilities

$$\begin{aligned}
\mathbb{E}[|E_{11}|] &= \frac{n_1(n_1-1)}{2} p_{11} \\
&= \frac{n_1(n_1-1)}{2} q_1, \\
\mathbb{E}[|E_{22}|] &= \frac{n_2(n_2-1)}{2} p_{22} \\
&= \frac{n_2(n_2-1)}{2} q_2, \\
\mathbb{E}[|E_{12}|] &= n_1 n_2 p_{12} \\
&= n_1 n_2 \sqrt{q_1 q_2}.
\end{aligned}$$

[3 marks]

We calculate the average degree of nodes of one type using a modification of the handshake theorem, *i.e.*, edges that connect between two nodes inside the type create a unit of degree at either end, whereas edges that connect to nodes outside the type create one unit of degree inside the type, so

$$\begin{aligned}
\bar{k}_1 &= \frac{2\mathbb{E}[|E_{11}|]}{n_1} + \frac{\mathbb{E}[|E_{12}|]}{n_1} \\
&= (n_1-1)q_1 + n_2\sqrt{q_1 q_2}.
\end{aligned}$$

Similarly

$$\bar{k}_2 = (n_2-1)q_2 + n_1\sqrt{q_1 q_2}.$$

[2 marks]

The average degree for a random node can be calculated by either (i) taking a weighted average of the two components above, *i.e.*,

$$\bar{k} = \frac{n_1\bar{k}_1 + n_2\bar{k}_2}{n_1+n_2},$$

or by taking the total number of edges, and re-using the handshake theorem to get

$$\bar{k} = \frac{2(\mathbb{E}[|E_{11}|] + \mathbb{E}[|E_{12}|] + \mathbb{E}[|E_{22}|])}{n_1+n_2} = \frac{n_1(n_1-1)q_1 + 2n_1n_2\sqrt{q_1q_2} + n_2(n_2-1)q_2}{n_1+n_2}.$$

[1 marks]

(b) Consider the number of triangles versus triples formed in this graph. In considering 3 nodes we might either have

- all three from the same type, or
- 2 from one type, and the third from the other.

When all three nodes come from the same type, then the clustering coefficient will be that of the Gilbert-Erdős-Rényi random network, *i.e.,* it will approach zero for a large network because the links are independent.

[1 marks]

When one node comes from a different type, we should be a little more careful, because now the edges are only conditionally independent.

[1 marks]

When one node comes from a different type, for instance, assume two nodes from Type 1, and the third from Type 2, then the probability of a probability of a triangle will be

$$P_{triangle} = Prob\{3\ edges\} = p_{11}p_{12}^2 = q_1^2 q_2,$$

and of a triple will be

$$P_{triple} = Prob\{3\ edges\} + Prob\{2\ edges\} = p_{11}p_{12}^2 + (1-p_{11})p_{12}^2 + 2(1-p_{12})p_{11}p_{12}.$$

Now for a large, sparse network, we can approximate $(1-p_{ij}) \simeq 1$, and so the clustering coefficient for these groups of 3 will be

$$\begin{aligned} C &\simeq \frac{3p_{11}p_{12}^2}{p_{11}p_{12}^2 + p_{12}^2 + 2p_{11}p_{12}} \\ &= \frac{3q_1^2 q_2}{q_1^2 q_2 + q_1 q_2 + 2q_1\sqrt{q_1 q_2}}. \end{aligned}$$

The important thing to note about this is that it has a cubic term in the numerator, and a quadratic term in the denominator. As the graph grows large, the q_i must go to zero to keep the average node degree constant (either using the above formula, or reasoning from first principles), and hence the cubic term will decrease more quickly than the quadratic, and hence this clustering term will also go to zero.

[1 marks]

Hence, the asymptotic clustering coefficient will be zero.

[1 marks]

4. (a) For an Eulerian cycle to exist, all nodes must be connected and have even degree.

[1 mark]

The degree of each node in a clique with n nodes is $n-1$. Hence,

(i) No.

(ii) Yes.

[2 marks]

(b) The diameter is the maximum eccentricity, or the longest (shortest) path in the network.

In this network, it is the length of the shortest path from opposite diagonals. For instance in the 2-dimensional case, we might go from the bottom left to the top right nodes. Each step to the right, or up takes one edge, so the distance of this path will be $2L$.

[1 mark]

(i) In general, we need to traverse L hops on each axis, and there are no short cuts, so the diameter will be dL.

[1 mark]

(ii) There will be $n=(L+1)^d$ nodes in the network. Thus

$$d = \frac{\log(n)}{\log(L+1)},$$

and so the diameter as a function of n is

$$D = L\frac{\log(n)}{\log(L+1)},$$

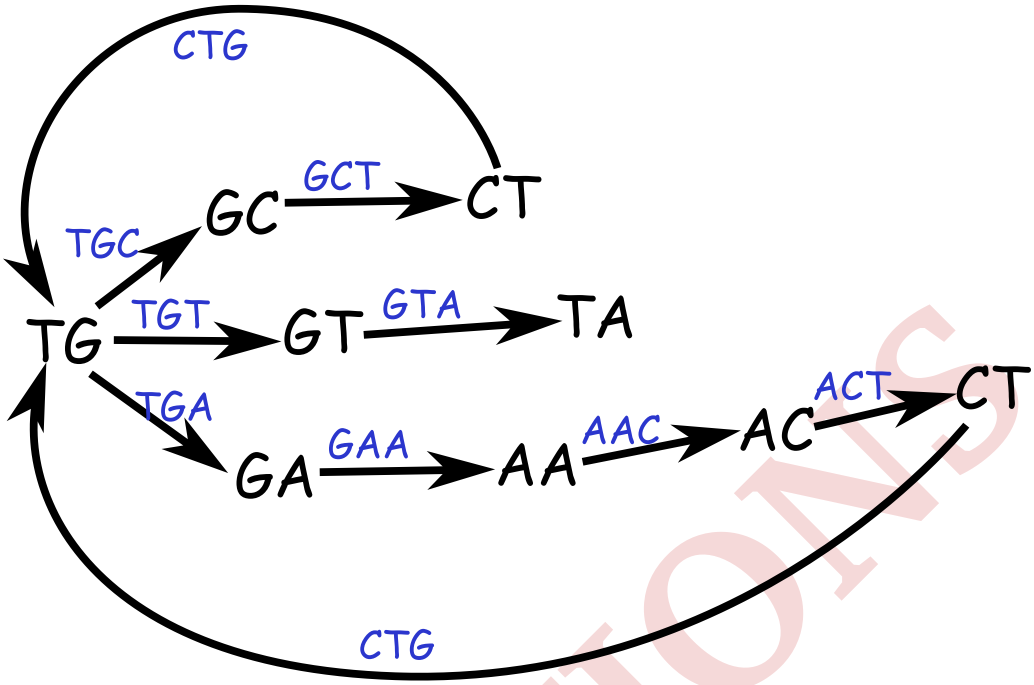
or in big-O notation (WRT to n) it is

$$D = O\big(\log(n)\big).$$

Either expression for D is acceptable.

[1 mark]

5. The figure below shows the de Brujin graph of the reads. Note that the edges are labelled by the reads, and the in-vertex of an edge is the prefix of the read, and the out-vertex is the suffix of the read. There are two edges labelled CTG, because this edge has multiplicity 2, all other edge-labels appear once.



[2 marks]

We can derive possible sequences from the graph by noting that

- the node TG has in-degree 2, but out-degree 3, so this must be the start point; and [1 mark]
- the node TA has in-degree 1, but out-degree 0, so this must be the final node. [1 mark]

However, there are two loops, and we don't know which order they occur in, so there are two possible sequences.

$$TGCTGAACTGTA$$
$$TGAACTGCTGTA$$

(the former was used in constructing the sequence).

[1 mark]

[1 mark]

6. There is no simple right or wrong answer to this question.

The following are a set of points to make relating the observed metrics to the model (the Gilbert-Erdős-Rényi random graph). I would expect students to make at least two of these points:

- The assortativity, while not large, is certainly larger than we would expect for a large Gilbert-Erdős-Rényi random graph (which should have assortativity $\simeq 0$). So the data exhibits more homophily (with respect to node degree) than the model.
- We would expect a large Gilbert-Erdős-Rényi random graph to have clustering coefficient close to 0. The observed coefficient is quite large at 0.621, so the data is much more clustered than the model.
- The average degree in the network is $\simeq 1,100$. If the network were Gilbert-Erdős-Rényi, then the degree of nodes would be binomial, here the probability of connection is approximately $0.05 \simeq 1,100/21,900$. The variance, therefore, of the node degree distribution would be (roughly) similar ($np(1-p)$ compared to np), and hence the standard deviation of node degree would be very roughly 30. However, the maximum node degree here is $\simeq 7,900$, which is very many standard deviations away from the mean, leading one to believe that the degree distribution is not binomial.

[2 marks]

All of these points show that the data is not consistent with the Gilbert-Erdős-Rényi random graph model.

[1 mark]

We can't, without some additional effort, determine exactly what effect these inconsistencies will have on the results of the paper, but it is quite conceivable that the results are completely wrong, because they are based on a model whose assumptions are inconsistent with the type of network that is actually being observed. It is very unlikely, given the large deviations between model and data, that there would be no consequences.

[1 mark]

Additional points that are relevant, but not necessary to make (and hard to see without access to the paper itself):

- In any approach such as this, there is a sampling effect, but the nature of the sampling, and its consequences are not explored.
- The network data in the data set must have been collected by one of the methods detailed in the paper, which notes that these methods are not 100% accurate. Hence, there is some circularity in the nature of the problem. This could have been finessed in the paper if the authors had examined the nature of the graph obtained from the inference process applied to the random network input.

[3 marks] will be allocated to the clarity and conciseness of the explanation of these details.

Please turn over for page 11

NUMBER OF QUESTIONS: 6
MARKS BREAKDOWN:

CORE	34
ADVANCED	16
TOTAL	50