PAPER CODE NO.	EXAMINER:	Prof S Maskell	
COMP 529	DEPARTMENT:	**CS**	TEL. NO: 44573

UNIVERSITY OF LIVERPOOL

FIRST SEMESTER EXAMINATIONS 2018/19

BIG DATA ANALYTICS

TIME ALLOWED: Two Hours

INSTRUCTIONS TO CANDIDATES

All candidates should answer **all three** questions

The numbers in the right hand margin represent an **approximate guide** to the marks available for that question (or part of a question). Total marks available are 100.

Additional Information:

None

UNIVERSITY OF LIVERPOOL

1)	a)	i)	What are the primary roles of the NameNode, Secondary NameNode and DataNode in an Hadoop Distributed File System (HDFS) cluster?	**3**
		ii)	Draw a diagram that shows the connectivity of the daemons that are associated with an HDFS cluster running on ten computers. Clearly show at least one of each of the following daemons: NameNode; Secondary NameNode; DataNode.	**6**
		iii)	Explain why an HDFS cluster, by default, stores multiple copies of each block of each file.	**2**
		iv)	An HDFS cluster with a default configuration has a blocksize of 64MB blocks. What is the total number of blocks needed to store three files if the three files' sizes are 6MB, 63MB and 620MB?	**3**
		v)	What are the two other daemons used extensively when using a Hadoop cluster to apply algorithms to data using MapReduce?	**2**
		vii)	Describe the input and output of the Shuffle-and-Sort component of a Hadoop cluster.	**4**
	b)	i)	Explain the role of tuples, a stream, a spout and a bolt in Storm.	**4**
		ii)	Name an alternative to Storm that can be used for Streaming analytics.	**1**
		iii)	Explain why the "heat wall" has given rise to an increase in the prevalence of multi-cored processors.	**3**
		iv)	Why are the Java language and Unix operating system used in the context of Storm?	**2**

Question 1 continues overleaf.

c) State the four Vs of Big Data. **4**

Total

34

2. Large ships transmit information that includes their position to ensure that other ships can avoid colliding with them. Systems exist that use a global network of satellites to collect the transmissions and create datasets relating to the trajectories of ships. Technologies exist to abstract each ship's trajectory into a sequence of a subset of the 5000 ports in the world that the ship has visited. There are multiple types of ship, e.g., ferry, tug, yacht, fishing boat and cargo ship. While some ships transmit their type, some do not. Your task is to use the behaviour of ships that do transmit their type to infer the likely type of other ships from such ships' behaviours.

a) i) What distributions should be used to describe: the count of how many times a yacht has visited each port during the last twelve months; your belief about the frequency that another ship that is known to be a yacht will visit each port next? **2**

ii) 100 ships are known to be ferries and 10 ships are known to be yachts. Draw a Bayesian Network, with plates, that describes the observed counts of how many times these ships have visited each port, the belief that results in the frequency that another ferry would visit each port next and the belief that results in the frequency that another yacht would visit each port next. **6**

iii) Assume that your belief associated with the frequency that ships visit the port of Dover next, x, is Beta distributed with parameters of α and β. Which of the following statements can be true: x=0.5; α=0.5; x=1.5; β=1.5? **4**

iv) Ferries visit Dover more often than yachts. Sketch the belief about the frequency with which ferries and yachts next visit Dover. Label the axes and clearly annotate each of two lines on the graph with the corresponding type of ship. Be sure to make clear any disparity in both the best estimate of the frequency and the confidence related to the two types of ship. Assume that the number of ports visited by a ferry per day is approximately the same as the number of ports visited by a yacht per day. **5**

Question 2 continues overleaf.

v) The Beta distribution is the conjugate prior for a Binomial likelihood. Will the order in which the ships visit ports affect your posterior belief? Why? **2**

vi) Data from an "unknown" ship exists: it is unknown whether the ship is a ferry or a yacht. Draw a Bayesian Network that describes the independence structure associated with the belief in the frequencies of port arrival for the two types of ship, the unknown type of the newly encountered ship and the observed count of port arrivals for this unknown ship. **5**

b) There are different types of fishing vessels (e.g., trawlers and longliners), but ships do not distinguish between these different types of fishing vessel in what they transmit. A large Bayesian Network can describe the statistical task of clustering these ships' trajectories. Using a table or otherwise, explain the relative merits of using Gibbs Sampling, Belief Propagation and Mean Field to perform this clustering operation. Be explicit about the algorithms' ability to exploit parallel computing hardware, the relative accuracy and any constraints on the Bayesian Network imposed by the use of each algorithm. **9**

Total 33

UNIVERSITY OF LIVERPOOL

3. Sensors on a gas turbine are used to monitor the operation of the gas turbine. You are part of a team developing tools to perform predictive maintenance of the gas turbine. While the sensors do provide measurements, the presence of measurement noise means that using these measurements as the input into an alerting system results in a prohibitively large number of false alarms. You have identified that it could be advantageous to pre-process the data using a Kalman filter.

a) There are three sensors on each gas turbine. Each sensor is assumed to monitor a scalar time-varying hidden state that evolves according to a nearly constant velocity model. How many dimensions do the measurement and hidden state each have? **2**

b) Measurements are received at a frequency of 10Hz. The process noise intensity is a parameter, q. What are the transition matrix and the process noise covariance matrix for the nearly constant velocity model used for each sensor? **5**

c) The measurements are recorded in units of Volts. Each sensor has a standard deviation of 1 Volt. What are the measurement matrix and the measurement noise covariance matrix in this case? **4**

d) Assume the transition matrix if F, the process noise covariance matrix is Q, the initial mean is μ and the initial covariance is Σ. Write expressions for the predicted mean and covariance in terms of F, Q, μ and Σ. **5**

e) Assume the measurement matrix is H, the measurement is y, the measurement noise covariance matrix is R the predicted mean is $\bar{\mu}$ and the predicted covariance is $\bar{\Sigma}$. Write expressions for the Kalman gain, the innovation and the updated mean. **8**

f) Concerns have been expressed about the computational expense associated with calculating the covariance in an online setting. Explain how to exploit the fact that the values of F, Q, H and R are constant to address these concerns. **2**

Question 3 continues overleaf.

g)	A company is monitoring five gas turbines. Draw a topology that describes each gas turbine generating data, Kalman filters processing each gas turbine's data and alerts being generated when two or more gas turbines exhibit unusual behaviour at the same time.	**5**
h)	Name an alternative algorithm to the Kalman filter that could be used in the topology if the model for the measurements was non-linear.	**1**
		Total 33