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

FIRST SEMESTER EXAMINATIONS 2016/17

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) You are setting up a Hadoop cluster. The cluster comprises 8 computers.

i) Which of the four Vs of Big Data does Hadoop primarily focus on? **1**

ii) What operating system would be used and why? **2**

iii) Why would you sensibly run the DataNode and TaskTracker daemons on the same computer? **2**

iv) Why would you sensibly avoid running the SecondaryNameNode and JobTracker Node daemons on the same computer? **3**

v) Draw a diagram showing how you would allocate daemons to computers. Show communications paths between the daemons. **8**

vi) What is the default number of replications of each block in a fully-distributed Hadoop cluster? **1**

vii) HDFS is now running on the cluster and is configured to have 64MB blocks (with default settings for the number of replications of each block). How would a file that is 321MB in size be stored in the cluster? **2**

viii) Explain how the NameNode would know if one of the DataNodes was unplugged while processing one of these blocks. **2**

ix) What is the name given to the configuration of Hadoop commonly used to debug MapReduce jobs? **1**

x) Why is java often used to describe MapReduce jobs? **1**

xi) Provide a brief summary of a generic MapReduce job in terms of the inputs, the outputs and their interrelationship for each of the map and reduce tasks. **6**

Question 1 continues overleaf.

b)	i)	In a Storm cluster, what name is given to the source of a stream of tuples?	**1**
	ii)	In a Storm cluster, what name is given to a component that takes (at least) one stream of tuples as input and outputs (at least) one stream of tuples?	**1**
	iii)	In a Storm cluster, what do the Nimbus and Zookeeper do?	**2**
	iv)	In a Storm cluster, what is a topology?	**1**
			Total 34

UNIVERSITY OF LIVERPOOL

2. You are building a system to analyse data related to crashes of autonomous cars for an insurance company.

a) Assuming that each individual car crash is an independent (instantaneous) random event and you wanted to use a conjugate prior, what distribution would you use for the likelihood (the number of car crashes that have happened) and for the posterior (your uncertainty related to the rate of occurrence of crashes)? **2**

b) The number of crashes is N_1 and the rate of crashes is λ_1. Write down an expression for $p(N_1)$ in terms of the prior, $p(\lambda_1)$, and likelihood, $p(N_1|\lambda_1)$. **2**

c) Using the answer from b) or otherwise, write down an expression for $p(\lambda_1|N_1)$ in terms of the prior, $p(\lambda_1)$, and likelihood, $p(N_1|\lambda_1)$ only. **2**

d) Autonomous cars have accumulated a total of t_1 driving hours. What is the average number of crashes by autonomous cars per hour? **1**

e) In total, humans have had $N_2 >> N_1$ crashes but have also accumulated a total of $t_2 >> t_1$ driving hours. Draw a graph with two lines, one describing the uncertainty related to the rate of crashes of autonomous cars and the other describing the uncertainty related to the rate of crashes of cars driven by humans. Label the axes and lines. Ensure that you annotate the graph with the average number of crashes for the two kinds of car and that it is clear how the uncertainties associated with these two estimates compare. Assume that the uncertainty is described using the conjugate distribution you named in the answer to part c). **6**

f) Explain why, even those autonomous cars might currently appear to be safer than cars driven by humans, there is a significant probability that they could eventually transpire to be less safe. **2**

Question 2 continues overleaf.

g) The posterior for *N* crashes during a time interval of *t*, can be approximated as being Normal with a mean of N/t and a variance of N/t^2. Data from the USA indicates that autonomous cars have had one crash in approximately 100 million hours of driving. Cars driven by humans have approximately 5 million crashes per year in approximately 5,000 billion hours of driving. Calculate the mean and standard deviation for both the posterior for autonomous cars and cars driven by vehicles. Comment on the ratio of the means relative to the ratio of the standard deviations. **3**

h) Where a car drives affects the chance of a car crash. Draw a Bayesian Network that describes the relationship between the number of car crashes, *N*, the rate of car crashes, λ, the location of the cars, *L*, and whether the car is autonomous, *A*. Assume that the location of the car is not directly dependent on whether the car is autonomous. **6**

i) To perform inference in a very much larger version of this graph involving many contributory factors relating to the risk of a car crash, it is proposed to use Gibbs sampling, Belief Propagation or Mean Field. What would be the relative advantages of each technique in terms of their ability to be parallelised, the number of iterations required and any restrictions on the graph necessary to use the techniques? A tabular answer is acceptable. **9**

Total 33

UNIVERSITY OF LIVERPOOL

3. a) The stream of images comprising a CCTV video of a person is being processed using a Hidden Markov Model (HMM). The HMM's states represent whether, at the time of the current image, the person is standing still, walking left, walking right, running left or running right. The video is recorded at 25Hz such that the time between images is 4 milliseconds.

i) The current state is x_t and the history of states up to time t is $x_{1:t}$. The model assumed for the dynamics of the state is a Markov model. What does this mean in terms of $p(x_t|x_{1:t-1})$ and $p(x_t|x_{t-1})$? **1**

ii) Populate a matrix with plausible values for the transition matrix. Assume that people must walk for at least 4 milliseconds before they run, must stand still for at least 4 milliseconds before changing direction and that people in every state are most likely to be in the same state 4 milliseconds later. **5**

iii) The measurements, y_t, extracted from the image at time t, provide a likelihood, $p(y_t|x_t)$ which is conditionally independent of the previous states. Does knowing x_{t-1} as well as x_t affect the calculated likelihood? **1**

iv) The output from the previous time-step is $p(x_{t-1}|y_{1:t-1})$. What is the minimum number of floating point numbers needed to be stored to describe this output? **1**

v) Express $p(x_t|y_{1:t-1})$ in terms of $p(x_{t-1}|y_{1:t-1})$ and $p(x_t|x_{t-1})$. **4**

vi) What matrix operation can be used to implement this expression? **1**

vii) Express $p(x_t|y_{1:t})$ in terms of $p(x_t|y_{1:t-1})$ and $p(y_t|x_t)$. **4**

viii) What name is often given to this equation relating $p(x_t|y_{1:t})$ and $p(y_t|x_t)$? **1**

xi) Draw the Bayesian Network for $p(x_{1:10}|y_{1:10})$. Do not use plates. Be careful to indicate which nodes are observed and unobserved. **6**

Question 3 continues overleaf.

b) i) Name three algorithms for sequential Bayesian estimation of a continuous state using nonlinear non-Gaussian models. For each algorithm, describe the approximation used. **6**

ii) Write an equation describing the likelihood model used by a Kalman filter when processing M-dimensional data to make inferences about an N-dimensional state. Define the size of any matrices used in the models in terms of M and N. **3**

Total 33