Quiz 5

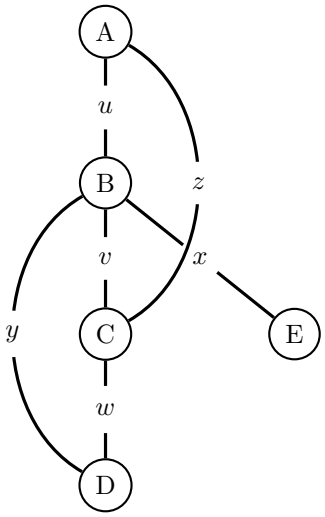
- There are 6 problems total worth 35 points as shown in each question.
- You must not communicate with other students during this test.
- No books, notes allowed.
- No other electronic device except calculators are allowed. You cannot use your mobile as calculators.
- This is a 45 minute exam.
- Do not turn this page until instructed to.
- There are several different versions of this exam.

1. Fill in your information:

Full Name: _____

NetID: _____

Zone 1



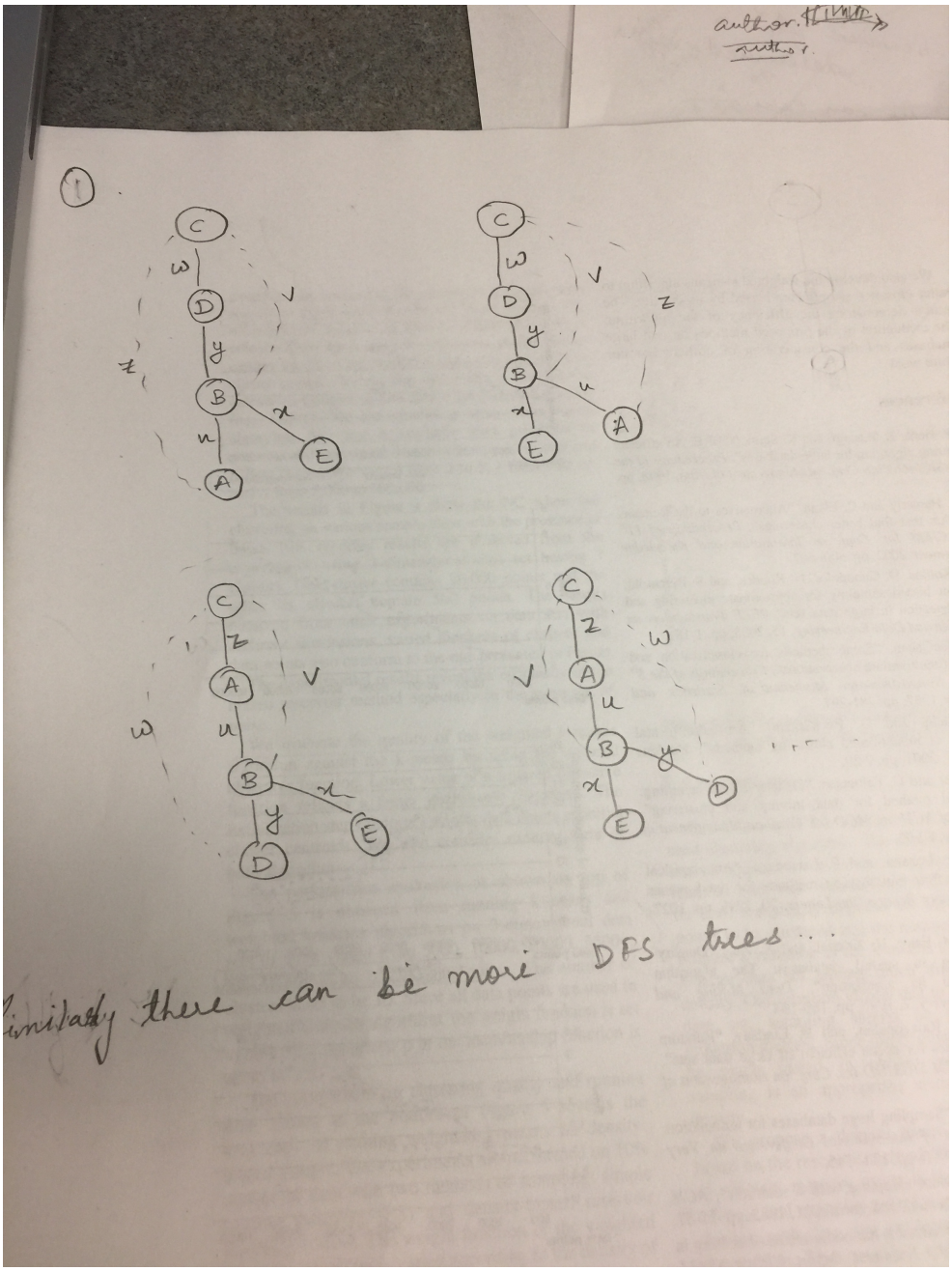
1/1. (5 points)

In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x, y, z).

Construct four different DFS search trees rooted at node C. Represent the DFS tree using solid lines for forward edges and dashed lines for back edges.

Hint: No need to consider lexicographical order while doing DFS traversal of graph.

Solution.



2/1. (5 points) For the following sequence database, identify all sequences that may have a prefix <**ab**>.

Sequence-id	Sequence
1	<a(abc)(ac)d(cf)>
2	<(ab)c(bc)(ae)>
3	<a(bc)(ab)(df)cb>
4	<a(ab)cbc>
5	<(ab)(ab)ccbc>

Solution. 3.

3/1. (5 points) Suppose the memory size of a sequence database is 16MB and the main memory available for computation is 32MB. Under such a scenario, explain the gain(or loss) of using PrefixSpan with pseudo-projections over PrefixSpan without pseudo-projections.

Hint : Efficiency of the two algorithms will be different.

Solution.

Since, the database can fit in the main memory, pseudo-projection saves time and space by avoiding physically copying postfixes.

4/1. (5 points)

The following example shows a candidate generation step in SPADE algorithm. Given in the tables below is the sequence-id and element-id of candidate sequences of length 1 (a, b, d, f). Complete the table for candidate sequences, <aa>, given the following information.

a		b		d		f	
SID	EID	SID	EID	SID	EID	SID	EID
1	2	1	2	1	1	1	3
1	3	1	3	1	4	1	4
1	4	2	1	4	1	2	1
2	1	3	1			3	1
3	1	4	2			4	2
4	3						

Solution.

aa

SID	EID(a)	EID(a)
1	2	3
1	3	4
1	2	4

5/1. (5 points) Shown below is a transaction dataset and item prices of each transaction. Assume each kind of item has one distinct positive price. Use the constrained FP growth algorithm to construct the FP tree that satisfy the requirements **min_sup = 2** and **min{S.price} < 1**. [question modified: there was typo in the question set because for sum the question becomes trivial.]

TID	Items (Price)
I1	1, 3, 4
I2	2, 3, 5
I3	1, 2, 3, 5
I4	2, 5

Solution. Check page 50 of advanced frequent mining slide.

6/1. (10 points) Consider the construction of a decision tree from the following training dataset. The training examples comprises predicting whether a person will buy-computer (Yes) or not (No) given two categorical attributes—age and education level. Number of possible age values is 8 and number of possible education levels is 4.

Age	Education	Buys computer
28	high	Yes
47	high	Yes
16	mid-	No
48	high	Yes
31	mid+	No
22	low	Yes
47	low	No
36	mid+	Yes
31	mid+	No
41	high	Yes

A. (4 points) Calculate the Information Gain obtained by splitting the decision tree by age and education respectively. Which of the two attributes is ideal for splitting.

Hint: $Gain(A) = Info(D) - Info_A(D)$;
$Info(D) = -\sum_i p_i log_2 p_i$;
$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$.

B. (4 points) Calculate the Gain Ratio obtained by splitting the decision tree by age and education respectively. Which of the two attributes is ideal for splitting.

Hint: $GainRatio(A) = Gain(A)/SplitInfo_A(D)$
$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2(\frac{|D_j|}{|D|})$

C. (2 point) Based on the above example, which of the two methods (Information Gain or Gain Ratio) is better for deciding the split. Please provide 1-2 sentence explanation along-with your answer.

Solution.

- $IG_{age} = .771$
  $IG_{education} = .495$

  Age is better attribute according to IG.

- $GR_{age} = .264$
  $GR_{education} = .268$

  Education is better attribute according to GR.

- Gain ratio is better. Information gain is biased towards attributes with a large number of values.
