Exercise 1

1. Consider the dataset in the following table, and compute the closed frequent itemsets with respect to the minimum support threshold $\theta = 3$. Show your work by first computing the frequent itemsets using the Apriori algorithm, and then finding the closed frequent itemsets.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

2. What is the definition of confidence of an association rule?

3. Why does the task of finding interesting association rules involve both a minimum support threshold and a minimum frequency threshold? What would happen if only exceeding a minimum confidence threshold was required for an association rule to be deemed significant?

Exercise 2

1. Describe the reservoir sampling scheme with an array of size $M$: the initialization (if any), and procedure at each time $t$ for each element $e_t$ on the stream.

2. What does it mean that the sample computed by reservoir sampling is a uniform sample?

3. Why does it make sense that the probability of modifying the sample decreases as $t$ grows?

4. Suppose that we run two instances of reservoir sampling independently in parallel on the same stream, each with an array of size $M/2$. Is the concatenation of the two arrays a uniform sample? Explain your answer.

Exercise 3

1. What are heavy-hitter vertices and heavy-hitter triangles?

2. Given a graph $G = (V, E)$, describe (in words or with detailed pseudo-code) how to find the heavy-hitter triangles. Discuss the time complexity of this approach.
Exercise 4

1. Describe (in words or with detailed pseudo-code) the preferential attachment model for graph evolution.

2. Describe the impact of the model on the properties of graphs generated according to it.

Exercise 5

1. Describe the HITS algorithm: settings, concepts of hub and authority, score definitions, and how they can be computed using the iterative method.

2. Consider the graph $G$ shown below, and compute the authority vector $a^{(1)}$ starting from the hub vector $h^{(0)} = (1/5, 1/5, 1/5, 1/5, 1/5)$, and then the hub vector $h^{(2)}$ using $a^{(1)}$. Assume that the additional constraint to avoid the homogeneity issue is to impose $\sum_{i=1}^{5} h_i^{(t)} = \sum_{i=1}^{5} a_i^{(t)} = 1$.

3. Describe how an attacker that can create infinite pages can increase the authority score of a target page.