Quiz 2 - Data Models Quiz

1. What is a possible pitfall of utilizing Excel as a way to manipulate small databases?

- **Excel does not enforce many principles of relational data models.**
- Excel is a user program and thus cannot run on a server.
- Excel does not allow algorithms for data manipulation.

2. What does the term "atomic" mean in the context of relational databases?

- Fixed schema of a particular database.
- A tuple that cannot be reduced.
- A column or row of data. Depends on the context.
- One unit of information that cannot be decomposed.

3. What is the Pareto-Optimality problem?

- Find the shortest path from source node to target node.
- **Find the best possible path given two or more optimization criteria where neither constraint can be fully optimized simultaneously.**
- Find the optimal path that requires going through specific nodes given by the user.

4. What constitutes a community within a graph?

- High density of nodes at a certain location.
- A neighborhood defined by an integer constant K around a specific node. All K+1 nodes belong in another community.
- **A dense amount of edge connections between nodes in a community and a few connections across communities.**
- Many anomalous neighborhoods within the same vicinity.

5. Why are trees useful for semi-structured data such as XML and JSON?

- Computers can easily visualize the data with a tree structure.
- It is not always the case that XML and JSON can be represented as trees.
- **Trees take advantage of the parent-child relationship of the data for easy navigation.**
- They are only useful for XML data as tree-like structure is apparent with tags. While JSON does not contain a tree-like structure as it contains arrays.

6. What is the general purpose of modeling data as vectors?

- Enables weighting of the query.
- The ability to normalize vectors allowing probability distributions.
- Enables image searching.
- **Results can be ordered by similarity using vector projection.**

7. For the following questions 7, 8, and 9, suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number.

Suppose we collect data month by month. Each month, we would have a batch of data containing the fields listed above. At the end of the year, we want to summarize our registrant activities for the entire year, so we would remove redundancies in our data by removing any records with duplicate account numbers from month to month. What type of operation do we use in this scenario?

- Join
- Not an Operation
- **Subsetting**
- Union

8. From the information given in question 7, what are the constraints, if any, which we have placed on the Account Number field for the end of year collection?

- Account should have at most n digits.
- If we had n duplicate Account Numbers then we will remove n-1 duplicate fields.
- There are no constraints.
- **Account Number should be unique.**

9. Suppose 100 people signup for our system and of the 100 people, 60 of them did not input an address. The system lists the values as NULL for these empty entries in the address field. Would this situation still have structure for our data?

- No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.
- **Yes the data has structure because we have placed a structural constraint on the data, thus the data will always have the originally defined structure.**