Exam — Answers — Business Intelligence and Process Modelling
Universiteit Leiden — Informatica & Economie, Friday June 2, 2017, 14:00–17:00

1. Instant view on current company performance leading to improved operational efficiency (1p), faster response to business change as a result of insight in possible trends and future business opportunities (1p) and easier communication/collaboration/information sharing with stakeholders (1p).

2a. Simpson's Paradox relates to the case when there appears to be a particular (e.g., positive) relation between variables in a dataset, where actually the data contains two or more groups for which an opposite relation (e.g., negative) is observed (3p).

2b. By means of visualization (e.g., using a scatter plot) the exact relation between the variables can be observed (1p), so that for example the presence of groups can be observed and a proper model to understand the relation between the variables can be chosen (1p).

3. The percentages do not sum to 100% (2p), the geographic representations wrongly suggests that certain races live in certain areas (2p), a two-dimensional visualization is used to represent one-dimensional data (2p) and the small groups are not visible, while one may be interested particularly in those (2p).

4. A data warehouse is already centered around subjects instead of applications and as such more focused on information acquisition (1p). It typically also contains (more) historic data to enable analysis over time (1p) and as opposed to a transaction system is nonvolatile, essential for consistent analysis (1p).

5. SMART: Specific, Measurable, Acceptable, Realistic and Time-sensitive (3p). Proper examples (2p).

6. Descriptive analytics deals with understanding and finding patterns in current data (1p), predictive analytics deals with predicting unknown or future aspects of the data (1p) and prescriptive analytics is about automatically making decisions based on analytics outcomes (1p).

7. One cannot be certain that there actually is a linear relationship between two variables (see for example Anscombe's Quartet). Inspecting the distribution of the data helps understand the true relationship.

8. In supervised outlier detection, we distinguish between items labeled as 'normal' or 'abnormal' using some classifier. Unsupervised outlier detection deals with finding instances that are least similar to all other instances, e.g., instances that fall into their own cluster as a result of applying a clustering algorithm. In semi-supervised outlier detection, first normal instances are characterized using a training data set, and then unlabeled examples are judged according to whether or not they fit the trained model.

9. Hierarchical clustering should be used if one expects a hierarchical structure (1p) or if one wants to a posteriori choose the level at which the clusters are defined (1p), whereas in k-means clustering the number of clusters is fixed in advance (1p). For large data, hierarchical clustering is infeasible given its quadratic time complexity (1p). k-means clustering is more affected by outliers (1p). Hierarchical clustering is deterministic, whereas k-means may suffer from random mean initialization (1p).

10. Occam's Razor is a principle which says that one should not increase the number of dimensions beyond what is necessary. PCA is a technique for dimensionality reduction which helps adhere to this principle.

11a. Indegree (centrality); the number of incoming links of a node is higher when many users like the node.

11b. Outdegree (centrality); the number of outgoing links of a node is higher when he/she likes more others.

12. The clustering coefficient is higher for networks with relatively large numbers of triangles. As we only consider heterosexual dating activity, triangles never form (the network is bipartite), and the clustering coefficient is always zero.

13. Although there is a giant component, there may still be smaller components not connected to the node where the rumor stats, causing the rumor to never reach its target (2p). There is also a temporal aspect. A rumor can only spread through links that occur sequentially in time (2p). Some users may go on one date before the rumor ever reaches what becomes their future monogamic partner.

14. Business intelligence is about analyzing and mining business data in order to make better business decisions attaining strategic goals. Process modelling is about formalizing processes within an organization into a model. Process mining considers doing this automatically, which could in a business context be seen as business process intelligence: analyzing and mining process data in order to (automatically) derive models that can be used to improve the business processes to meet strategic goals.

15. Petri nets have formal verification. Petri nets can better model concurrent processes. BPMN has more primitive building blocks, so that complicated processes can be written in a more compact way. Subprocesses can more easily be embedded in BPMN. BPMN is more readable for non-experts.

16. A new place with three incoming arcs from 'g' and one outgoing arc to 'h' needs to be added.

17. Multiple answers are possible. 3p for the frogs and their jumping/swimming/hopping behavior. 2p for incorporating the princess and 3p for properly counting each third frog. See Figure 1 (left).

18a. 1) Fitness (the discovered model should allow for the behavior seen in the event log), 2) Precision (the discovered model should not allow for behavior completely unrelated to what was seen in the event log, 3) Generalization (the discovered model should generalize the example behavior seen in the event log) and 4) Simplicity (the discovered model should be as simple as possible).

18b. Fitness: yes (everything fits), precision: no (much more behavior than in a real-world process is possible), generalization: yes (although likely too much) and simplicity: yes (likely too simple, though).

19. Explain how k-fold cross-validation works (2p), and mention how it related to process discovery: training a process model on event logs (2p).

20.
```
#  -> #  -> #  #
<- #  || || || ->
#  || #  -> <- #
<- || <- #  -> ->
#  || -> <- #  #
#  <- #  <- #  #
```

$T_L = \{a, e, c, d, b, f\}$, $T_I = \{a\}$, $T_O = \{f\}$
$X_L = \{(\{a\}, \{b\}), (\{a\}, \{d\}), (\{b\}, \{f\}), (\{c\}, \{d\}), (\{d\}, \{e\}), (\{d\}, \{f\}), (\{e\}, \{c\}), (\{a, c\}, \{d\}), (\{d\}, \{e, f\})\}$
$Y_L = \{(\{a\}, \{b\}), (\{b\}, \{f\}), (\{e\}, \{c\}), (\{a, c\}, \{d\}), (\{d\}, \{e, f\})\}$
P_L is exactly the set of places between the transitions in Y_L plus a start and end place.
Then deriving F_L is somewhat trivial. Next, $\alpha(L) = (P_L, T_L, F_L)$. See Figure 1 (right).

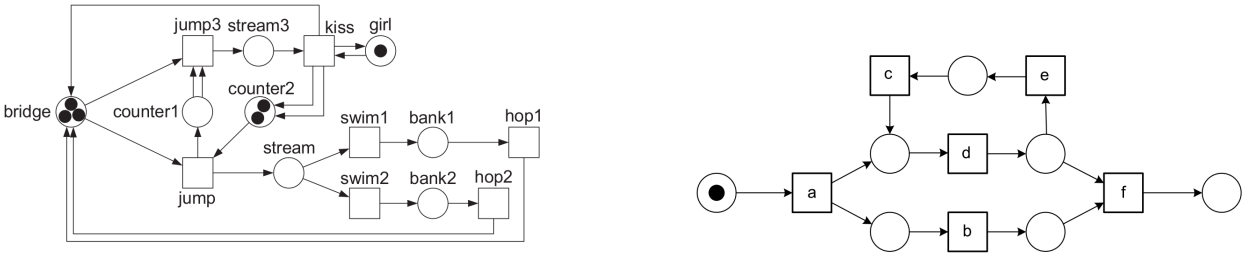

Figure 1: Possible model for question 17 (left) and the final Petri net for question 20 (right)