Exam

Data Science and Process Modelling

Universiteit Leiden — Informatica & Economie

Wednesday May 29, 2019, 14:00–17:00

This exam consists of **20 questions** divided over 4 pages. Answers can be in **Dutch or English**. Always give **precise, well-motivated and to-the-point answers**. Write down any nontrivial assumptions. The number of points awarded for each perfectly answered question is listed in front of the question, and sums to **100 points**. Your grade is computed by dividing the number of points by 10. Good luck!

(12p) Visual Analytics & Event Data

1. (3p) What do the concepts of proximity, connectedness and similarity refer to in the context of data visualization?
2. (3p) Explain how Simpson's Paradox highlights the importance of data visualization.
3. (2p) How does "knowledge-driven" detection of anomalies work?
   Hint: examples of applying this technique were discussed in the guest lecture about Eventpad.
4. (4p) Describe at least two things that are wrong about the visualization shown in Figure 1.

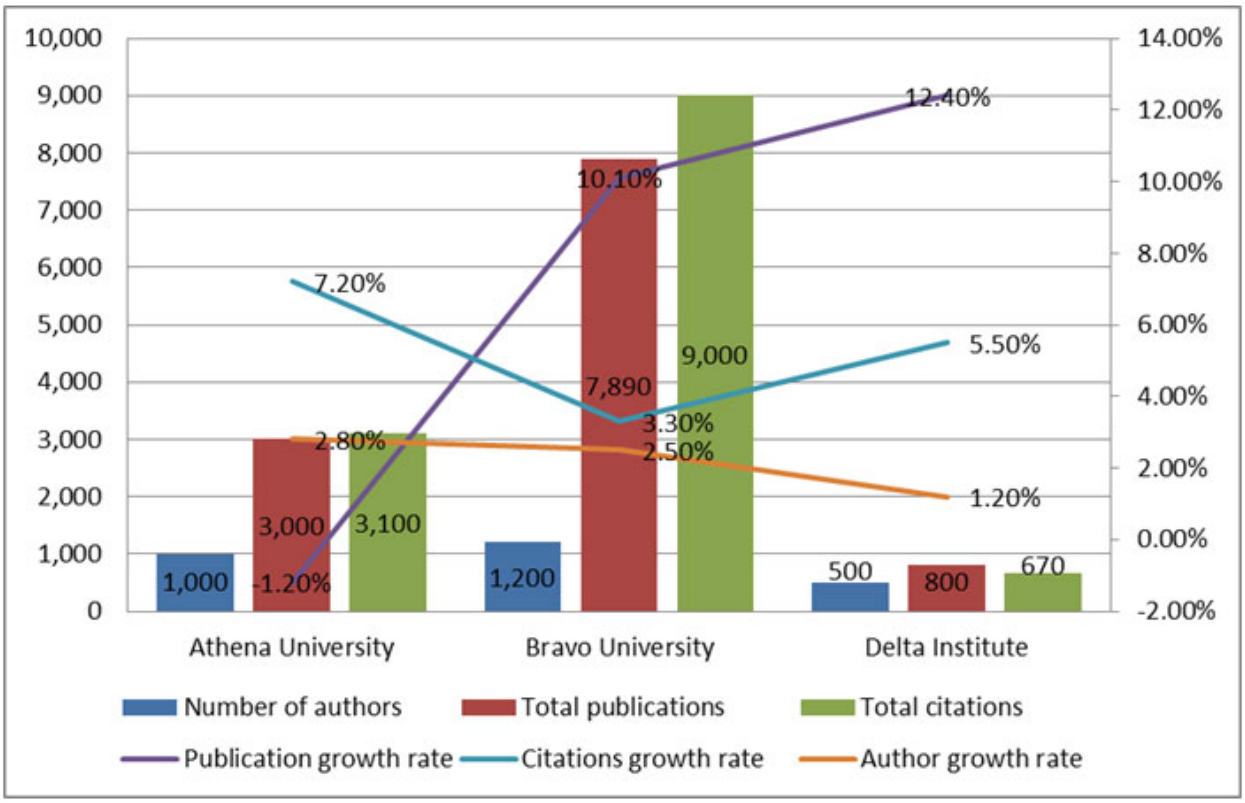

Figure 1: A visualization of scientific performance measures.

(38p) Data Science

5. (6p) Analyzing modern 'big data' does not only mean coping with the large volume of data, but typically also refers to at least three more aspects that make the analysis nontrivial. Name these three aspects, and explain them briefly in the context of medical patient data.

6. (4p) What is the difference between data-driven research and hypothesis-driven research? Give a short example of both.

7. (4p) To understand the relationship between two numerical attributes of a tabular dataset, one can compute the value of the Pearson correlation between the two attributes. Give two reasons why this value alone does not say everything about the relation between the two variables.

8. (4p) Scientists have recently estimated that there are approximately 8.7 million species on Earth. One of these species can be described by attributes such as its size, color, type of food, habitat, region of origin, etc. Give an advantage and a disadvantage of applying hierarchical clustering to cluster these species into meaningful groups.

9. (4p) Why can the XOR-problem never be learned by a perceptron?

10. (4p) What is the difference between local encoding and distributed encoding ("locaal coderen" and "gedistribueerd coderen"), in the context of representing the input or output of a neural network?

Case: Instagram network. Instagram is a large social networking platform for users across the globe. On Instagram, users can follow other users to see their updates and photos. These follower relationships (links) create the so-called 'Instagram network' of directed links between nodes (users). Apart from regular users, there are a number of so-called 'verified' users; typically celebrities and other prominent people.

11. (4p) What do the two measures of indegree and outdegree indicate in the considered Instagram network?

12. (4p) It is often said that celebrities in social networks can be distinguished from regular users based on their position in the network. Both celebrities and regular users can have a high degree, so that does not say much. But, we can look at the number of connections between a node's neighbors. More specifically, it is said that celebrities draw their links from seemingly random users all over the world. In contrast, regular users are frequently connected to users that together form a more tightly connected group of users.
    Which node measure can be used to assess the connectivity of a node's neighbors, and thus whether the node might be a celebrity? What does this measure indicate in general?

13. (4p) Name two centrality measures other than (in/out)degree centrality applicable to the Instagram network, and explain which aspect of node importance these measures capture.

Figure 2: Logo of the Instagram platform. This exam was not sponsored by any corporate entities.

(50p) Process Modelling

14. (6p) Explain the concepts of 'concurrency' and 'mutual exclusion' in the context of process modelling.

15. (10p) Consider the grading process of a course at a university, for which a student can obtain a sufficient ("voldoende") or insufficient ("onvoldoende") grade. This grade is submitted to the student administration. We are interested in the situation of one student. The student's grade consists of one exam and three separate assignments, that can be handed in in any order. All parts of the course can be retaken once, if the result of that part is insufficient.
    There can be different situations:

    - If the three assignments and the exam all result in a sufficient grade, then this final sufficient grade is sent to the student administration.
    - If the assignments are not sufficient, but the exam is sufficient, then no grade is submitted, until the assignments are completed.
    - If the grade for the exam is insufficient, then regardless of the result of the assignments, the insufficient grade is sent to the student administration (but, if it is the first exam try, a retake can still result in a sufficient grade).

    Model this situation as a Petri net.

16. (6p) Describe how encapsulation is achieved in BPMN, and how it can be realized in Petri nets. Describe a notable difference between the two.

17. (6p) Give the reachability graph of the Petri net in Figure 3.

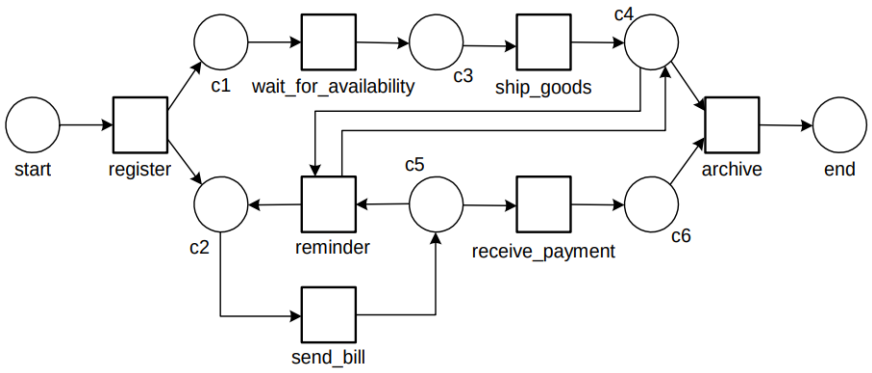

Figure 3: A Petri net representing order processing.

18. (4p) Name and explain four criteria for judging the quality of a discovered process model.

19. (6p) Explain how the problems of incompleteness and noise play a role in the creation of a process model.

20. (12p) Apply the α-algorithm to derive the footprint and final corresponding Petri net of event log L.

$$L = [\langle a, b, d, g\rangle, \langle a, d, b, g\rangle, \langle a, c, d, e, f, g\rangle, \langle a, c, e, d, f, g\rangle, \langle a, c, e, f, d, g\rangle, \langle a, c, f, e, d, g\rangle,$$
$$\langle a, c, d, f, e, g\rangle, \langle a, c, f, d, e, g\rangle, \langle a, c, d, e, f, g\rangle, \langle a, d, c, e, f, g\rangle, \langle a, d, c, f, e, g\rangle]$$

    Explain your steps in detail. After the derivation of the places, you can immediately draw the net; there is no need to write out all arcs in formal notation.

End of exam. Please do not forget to fill in the evaluation form!