Research School of Computer Science, Australian National University

COMP3420, Advanced Databases and Data Mining

Assignment 1

Due: Thursday 31 March 2016, 5 pm

Front Matters:

- There are 7 problems in this assignment. There are 6 written questions and 1 data analysis question. A calculator or computer may be needed for the written questions, the data analysis question can be done in any tool of your choice – spreadsheet, Rattle, R, Python, or other programming languages.
- This assignment will be graded out of a total of 100 points, it is worth 20% of the final marks.
- This assignment must be submitted in electronic form via Wattle. We only accept PDF documents (in one single file, typesetting or scanned – with typset preferred), no other formats such as Word or OpenOffice documents, etc. On Wattle, there will be a link called "Assignment 1 Submission" – follow instructions there to upload your assignment. It is your responsibility to ensure that the uploaded PDF file is clearly legible and printable.
- Please clearly put your **student ID** at the top of the first page of your submission. **Do not** put your name on the assignment sheet since grading will be blind.
- Note that you are required to show all your major working steps for all calculation questions. In other words, if you just write down the final result for a question, you would not receive credit for that question.
- Note that the questions are sorted by theme and example, and not sorted according to their difficulty.
- Late policy: we adhere to the standard ANU policy of special considerations. Late penalties are assessed as 5% of obtained mark per every 24 hours late, or part thereof. The **submission cut-off is at 4 days**, i.e. an assignment submitted more than four days late after the submission deadline will receive no marks.
- Exceptions: Should there be a medical condition or other unfortunate circumstances beyond the student's control, it is the student's responsibility to get in touch with the lecturers *before* the original deadline to agree on an alternative arrangement, and be prepared to show proof.

Question 1 [20 marks] Data Mining in an Online Game

In this problem we examine data from *Magic League*, an online *role-playing game* (RPG), where each player assumes a virtual identity, aims to advance in a virtual landscape, and earn scores by defeating virtual monsters. Even for the most prolific geeks, spending hours shooting virtual monsters can get lonely. Therefore, game developers have introduced online multiplayer options, in which gamers register accounts, they chat with other users, they form online alliances and clans, and they share virtual quality moments together fending minions.

Your mission is to study the user and games data and help improve the social aspect of undead shooting. The user data is stored in the following tables:

1. **Accounts**. These include: information about each user such as their screen name, chosen character, date they sign up, best scores, unlocked achievements, held artifacts, experience level and hit points.
2. **Alliances**. This database includes information about user alliance composition, recording for each alliance the id of its members.
3. **User Interactions**. This table contains information about user-to-user interactions. The information include: the time of the interaction, id of the two users involved in the interaction, the type and attribute of the interaction (e.g., artifact exchange, fight, or chat).
4. **Battle Sessions**. This contains information about each game session. Including: the time that a user logs in; the time the session starts and ends; each user's actions, losses and achievements; and the outcome of the battle.

Your overall task is to build a data warehouse from these databases to analyse user behaviour, with the goal of building a game that is even more fun to play.

(a) A gamer, Tintin, logs into his account and together with a few friends from his alliance started to conquer a forest occupied by monsters. In this battle, Tintin was bitten by a unicorn and temporarily lost his ability to fight. He recovered by drinking a life potion offered by an ally, Pinocchio. The battle ended with the alliance taking control of the forest. Which databases tables (among 1–4 above) are changed during this session? And if changed, are entries added or updated? Please indicate your answer by circling the appropriate option for each DB below. **[4 marks]**

DB(1) Accounts: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(2) Alliances: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(3) User interactions: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(4) Battle sessions: **Changed / Unchanged**; (if changed) **Added / Updated**

(b) Tintin and Pinocchio's team needs to assign a team member to fight a new monster, *Basilisk*, in the *reptile* category. You need to supply a function that helps them decide who has had the most number of wins against *reptiles*. Which database(s) do you need to use to get the information? **[2 marks]**

(c) What is metadata of a database? Please provide a brief description. Also give two examples of metadata about the **Battle Sessions** database above. **[4 marks]**

(d) You are to generate a data warehouse, containing the average number of battles users participated in, and the average number of logins they made of in May 2014; tabulated by the alliances they are in, and the year they signed up for the game (2011, 2012, 2013). Draw a star schema of this table. Clearly annotate where each measure and each dimension come from in databases (1–4). **[4 marks]**

(e) You analysed the game achievements versus user interactions for a number of prominent alliances in May 2014. Compute the average size of alliances in this sample. **[1 mark]**

Alliance Name	Size (# Members)	Number of Monsters Defeated per Member	Chat Messages Exchanged per Member
Justice League	7	150	2.5
Gryffindor	150	4.2	300
Dumbledore's Army	12	83.0	18.5
Fellowship of the Ring	6	0.0	260
Southern Airbenders	45	23.5	25

(f) In the game achievements table in the previous part, how does the average number of defeated monsters related to the average number of chat messages exchanged? Answer this question by computing the Pearson correlation coefficient of these two quantities. Show your workings. **[3 marks]**

(g) Assuming the data in part (e) is collected correctly, what do you think *could* lead to the observed behaviour in each alliance? Provide one example reason. **[2 marks]**

Question 2 [10 marks] Understanding User Data

Consider the following set of 6 game users.

ID	Name	Gender	Age	Skill Level	Motto
u001	Harry	Male	17yrs	Black	"Never tickle a sleeping dragon"
u002	Hermione	Female	16.5yrs	Brown	"When in doubt, go to the library"
u003	Katara	Female	15yrs	Black	"Never turn my back on people who need me"
u012	Frodo	Male	16yrs	Green	"One ring to rule them all"
u011	Superman	Male	25yrs	Black	"An acceptable face of invading realities"
u066	Astro	Male	15mo	Doggy	"Happiness is a warm puppy"

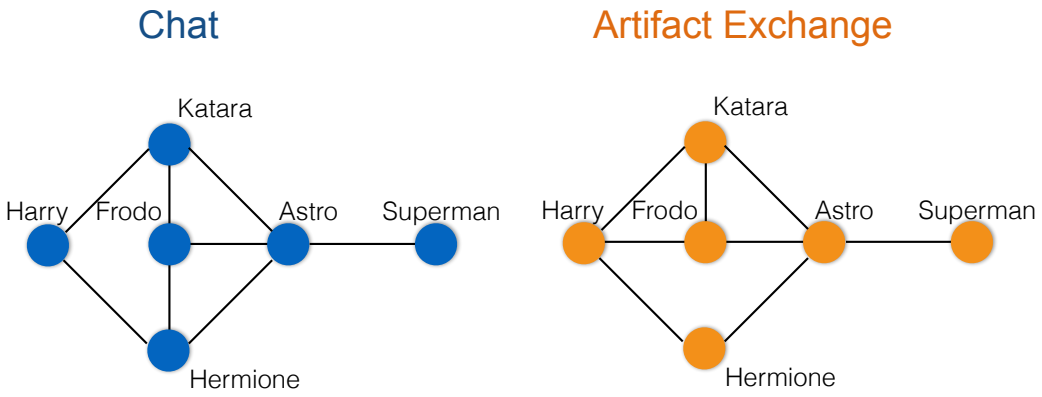
(a) What is the number of datum n and the number of attributes p in this table? **[2 marks]**

(b) Among the p different attributes, identify one binary and one numeric attribute. **[2 marks]**

(c) You need to apply a data mining algorithm which only accepts binary attributes on this user profile dataset. Explain how the age and skill level variables can be transformed into a binary attribute, or a set of binary attributes **without loosing any information contained the original dataset**. Write out the transformed binary attributes for the skill level attribute for each user. **[6 marks]**

Question 3 [12 marks] Constructing User Graphs

We construct two graphs among these six users, by making an (unweighted) edge between two users when they have exchanged at least 5 messages in May 2014 (on the left, the *Chat* graph); and when they have exchanged at least 5 virtual artifacts in May 2014 (on the right, the *Artifact Exchange* graph).



(a) How many edges are there in each graph? **[2 marks]**

(b) Which node(s) have the highest degree in each graph? **[2 marks]**

(c) Which node(s) have the second highest degree in each graph? **[2 marks]**

(d) What is the closeness centrality of Frodo in each graph? **[2 marks]**

(e) What is the (un-normalized) betweenness centrality of Katara in the *Chat* graph? i.e. the number of shortest paths from all users to all others users that pass through Katara. Is this the same with her betweenness centrality in the *Artifact Exchange* graph, why or why not? **[4 marks]**

Question 4 [8 marks] Data cubes and OLAP

The *Magic League* game provides a feature for users to build virtual pets, take them along as battle companions, or give to each other as gifts.

1. `Species` ten possible values: Dog, Cat, Pig, Lizard, Horse, Stag, Otter, Swan, Hare, Phoenix.
2. `Gender` two possible values: Male/Female.
3. `Color` seven different values.
4. `Size` five different values.
5. `Intelligence`, five different values.

You are to study users pet-keeping behaviour.

(a) The first task is to construct a data cube. How many cells are in the base cuboid? **[3 marks]**
(2 marks for working, 1 mark for the correct final answer)

(b) How many cells are there in total if you were to compute all cuboids? **[3 marks]**

(c) One summary measure in the datacube is the number of pets. Given a cuboid with dimensions `Species`, `Colour` and `Intelligence`, what OLAP operations do you use to get the number of pets that is *purple*, and tabulated by their `Intelligence`? **[2 marks]**

Question 5 [15 marks] Structure of a Network

Consider the set of 18 Web pages drawn in the following figure, whose links forming a directed graph.

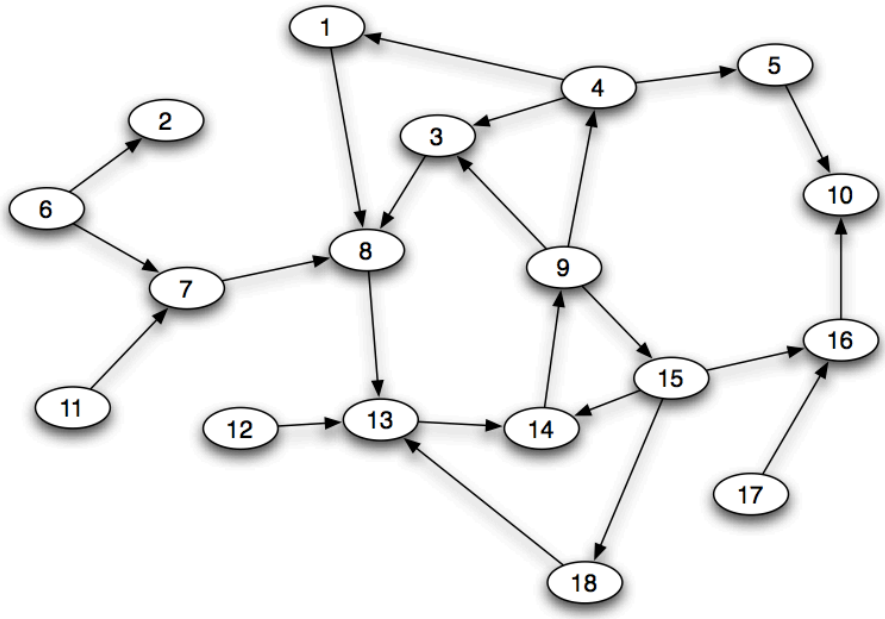


Figure 1: A directed network of 18 web pages.

(a) Which nodes constitute the largest strongly connected component (SCC) in this graph? Taking this as the giant SCC, which nodes then belong to the sets IN and OUT as defined in the lectures? Which nodes belong to the tendrils of the graph? Explain all of your answers. **[9 marks]**

(b) As new links are created and old ones are removed among an existing set of Web pages, the pages move between different parts of the bow-tie structure.

Name an edge you could add or delete from the graph in the above figure so as to increase the size of the largest strongly connected component. Explain why you named this edge. **[3 marks]**

(c) Name an edge you could add or delete from the graph in the above figure so as to increase the size of the set IN. Explain why you named this edge. **[3 marks]**

Question 6 [15 marks] Chi-Square Test

Hogwarts owlry keeps a large number of owls with varying magic capacity. We examine their *feather colour* - black or white; and *beak colour* - red or yellow, along with a critical magic property: *ability to locate the recipient* - strong or weak. The table below contains the number of owls that possess two qualities simultaneously, e.g. there are 10 black-feathered owls that have strong localization ability.

	black feather	white feather	red beak	yellow beak
weak localization	45	30	60	15
strong localization	10	15	20	5

(a) How many owls have black feather? white feather? How about red or yellow beak? **[4 marks]**

(b) Is the magic localization ability correlated with feather colour or beak colour? Which feather or beak colour seems to produce highly capable owls? Answer this question by manually computing χ^2 tests on feather/beak colours and localization ability. Show all of your workings. **[7 marks]**

(c) If there is a third attribute, having *sulphur crest*, found to be highly correlated with strong localization ability in owls, with $\chi^2 = 20$. Is it correct to say that *sulphur crest* causes improved localization ability? Why or why not? **[4 marks]**

Question 7 [20 marks] Hands-on Analysis of a Real-World Dataset

Take the UCI Energy efficiency dataset `http://archive.ics.uci.edu/ml/datasets/Energy+efficiency`, complete the following analysis and compute the designated metrics.

(a) Draw a boxplot of dimension $y1$: `heating load`, annotate all key landmarks on the box plot. What is the median of $y1$, what is the mean of $y1$, what are the values of $Q1$ and $Q3$, how large is the inter-quartile range, are there any outliers?

[8 marks]

(b) Draw a scatter plot of dimension $Y1$: `heating load` vs $Y2$: `cooling load`. What is the minimum and maximum of $Y1$ and $Y2$, respectively? Are $Y1$ and $Y2$ positively correlated, negatively correlated, or appear to be un-correlated?

[5 marks]

(c) Compute the Pearson correlation of the 8 attributes $X1 \ldots X8$ with $Y1$ `heating load`.

[4 marks]

(d) Rank the attributes with respect to their strength of correlation with $y2$ `cooling load` – from the least to the most correlated.

[3 marks]