University of Glasgow Hackathon Challenge Question

Huw Evans
SAS, Inc

relationships.txt is an 8GB file that contains records of the following form:

```
7084518844172374795
{
  DISLIKES 8811886287792076713
  FRIEND_OF 3793997454376268553
  KNOWS 9073105514090943087
  MARRIED_TO 6850878856176462406
}
```

The above record describes a person and their relationships with other people. The above record shows that someone with a person id of '7084518844172374795' has four relationships with others. This person dislikes another person whose id ends in '6713', they are friends with a person whose id ends in '8553', and so on.

The challenge is to:

1. Read in relationships.txt and build an internal data structure that will support the querying of the data set
2. Write code to query the data to answer the following questions:
    a. How many people are there?
    b. What is the average number of relationships for people with odd and even person id values?
    c. How many people have relationships with themselves?
    A person has a relationship with themself if a person's id is mentioned in their own record.
    d. Between two people, how many FRIEND_OF relationships are returned?
    A FRIEND_OF relationship is returned if person A is FRIEND_OF person B and person B is FRIEND_OF person A
    e. Which person is disliked the most?
    f. Visualise the relationship graph with a user interface via a scrolling two-dimensional space similar to Google Maps
    g. Which person has the longest cycle of FRIEND_OF relationships that eventually leads back to the same person?

Points to Note

The person id is a positive 64-bit integer.
It is possible that the person ids are not unique.
There are five relationship types: DISLIKES, FRIEND_OF, KNOWS, MARRIED_TO, HAS_DATED.
Consider how the internal data structure and the types it is built from perform as more and more data is loaded into them.
A visualization system for question f. would take quite a while to write from scratch. Consider how to speed up the writing of this.
Think about what question g. is asking for before you write any code. Drawing diagrams may help you understand what is required.

Evaluation Criteria

Answers for 2a. to 2g. will be judged on the following criteria:

1. Correct values for 2a. to 2e.
2. 2f. will be judged in terms of visualization of the largest part of the dataset (or all of it).
3. 2g. will be judged in terms of speed of calculation. The correct answer calculated in the shortest amount of time will win.
4. In the event of a tie-break, speed of calculation and visualization time for 2a. to 2g. will be used to determine a winner.