1 Let us suppose that the distance function is defined for the data set described purely by categorical features. Does it necessarily mean that each feature element of the data set may be ordered?

Originally, categorical data is not ordered and can not be ordered by itself, or at least it does not make sense. If we can compute a distance function (and not a similarity function !), that means that we associated numerical values to the data. In this case, we can order the data with those values, technically, but it does not really makes sense as we are the one to assign a value to the data.

2 Referring to the computation process of Dynamic Time Wrapping distance between sequences, how should the case "repeat neither" be interpreted?

The case *repeat neither* in the case where $DTW(i-1, j-1)$ is the minimum between $DTW(i-1, j)$, $DTW(i, j-1)$ and $DTW(i-1, j-1)$. The principle of DTW is to map point by point to sequences of data, that can have different lengths. In a simplified way, if we have two sequences X and Y that have two different length and we want to match one point to another, we will sometimes need to repeat some point so both can match, and this is what the *min* part is representing. When the minimal value is $DTW(i-1, j-1)$, that mean that none of the point need to be repeated in order to connect points from both sequences; it is a match between those points.

3 Describe the main difficulty in designing distance functions for the data sets which features are of a mixed type (categorical and numeric).

The main difficulty is that we need to arbitrary define the order for categorical data, which is not really adapted to distance function (similarity function is a better idea). Therefore, depending on the choice we make we can influence the output, and sometimes compromise our results.

4 Referring to the time-series data mining, which distance measure is more informative, DTW or Motif based?

DTW would gives more information as it can be applied on sequences of different length, and can also gives us information even if the pattern are not the same but are similar. Motif based distance will only count motif that are the same length, which can be useful in special case. We can also compute motif of different length if we use DTW as the distance function for the distance based motif, therefore, using both at the same time would give us the most information.

5 Describe the outlier in the context of time series data mining problems.

An outlier in a time series could be an abrupt and unpredictable change, during a relatively short amount of time. Either one point is an outlier and is far from the forecast, or whole pattern of point can be outside of the general tendencies without each one being an outlier considering around point.

For example, some events, such as terrorist attack, can cause stock market to have be unpredictable for a time. There is also the case when a heart has a weird beat: considering point to point, we can not see an anomaly but globally, the tendency is not the same.

6 How could one change the reservoir sampling algorithm to guarantee a higher presence of the observation point from a specific time period? For example, the data points between current time minus five hours and current time minus eight hours.

Modifying the bias function that gives the probability of the point to be part of the sample could allows us to change the quantity of point from a specific time period. For example, giving an higher probability for point in current time minus five hours and current time minus eight hours. That does not mean we will only have data from that time period, but only a higher presence of them.

7 Describe in your own words the phenomenon of concept drift.

The concept drift describe the fact that the data distribution can vary in time. That means that properties of the data also evolve during time : correlations, cluster distributions, ... This can influence negatively the accuracy of the prediction : for example, the prediction of the number of customer for a restaurant will not be the same on Monday evening than on Saturday evening, or on Tuesday evening and at lunch. I would say this in a way, kind of the effect of the "time context" on the data : the time going by affect the data and the data prediction because the context of it changes.

The decay-based reservoir sampling can help handle concept drift: the more recent data is considered more important, and the older the data get, the less important it is, which mean the prediction is based on the tendency of the most recent data.

8 How the super problems of data mining (classification, clustering, association pattern mining, outlier analysis) are translated to the cases of text data mining.

The super problems are transferred in text data mining using diverse modification on the algorithm and data-preprocessing on the text dataset. For example, representing words by vectors with a signification in a way that we can apply usual algorithm for clustering and classification. Notably, the tf-idf representation use the frequency of the word and the inverse document frequency. The distance function to use is also different, and we can also use similarity instead of a distance function.

9 What is the main disadvantage of the probabilistic approach in text data mining?

There is different ways to approach in a probabilistic way text data mining. First is PLSA, which can be really useful for a little corpus, but can be really slow with a lot of document as the number of parameters grows linearly with the size of the corpus. Also, this approach can stick to close to the training data and gives back results not applicable for another case and for the test data (overfitting).

Another known algorithm, we already studied with numerical data, was the EM algorithm, with a soft clustering practice. With soft clustering, the datapoint is not assigned to a cluster

but has a probability to belong to each cluster, which mean that it can have almost the same probability for 2 clusters for example, and also allows cluster to overlaps.

10 What is the main property of the distance measures used for graph data mining?

There is two type of distances in graph datamining : between graph and between nodes. The main properties used to compute the distance between node is the shortest path between those node, however, there might be no path between those nodes, which can be considerate as an infinite distance in a way. Between two graph, the main properties is common subgraph, and computing the number of common subgraph can be a way to measures the distance between two graph, or also using topological descriptors.

11 Explain the meaning of topological descriptors?

Topological descriptors use quantitative measure from structural graph as dimensions. They are indexes used to describe the structure and the properties of the graph. Some topological descriptor are node specific, as Morgan Index (number of other node reachable within a distance k), or graph specific, as Weiner Index (sum of the pairwise shortest path). In some domain, as in chemistry (crystallography) topological descriptors help studying the structure of matter.

12 Do neighbourhood measures belong to the set of topological descriptors?

Neighbourhoods measures can belong is the set of topological descriptors, as node specific descriptor. It is a way to have an insight of the global structure of the graph, and how connected are the nodes inside the graph, so in a way, how dense is the graph.

13 Among the node properties in social networks, which three are the most informative?

Social network are directed graph usually (relation are not always reciprocate). I would say that the degree prestige, the gregariousness of the node, and proximity prestige. The degree prestige allows to know how much other nodes follow the node, and represent the popularity of the node inside the network. The gregariousness does not gives the same information

about the node and gives us information about how much the node is searching for relations. The proximity prestige then allows us to know how central the node is, how much influence it has.

14 Regarding privacy preservation, which stages of data mining workflow are affected the most.

The most affected stage are the data collection and the publication, and the output of the algorithm. During data collection, random noise can be added in order to "blur" user data, which mean needing to work with the data distribution instead of the data record.

During publication, it is also needed to remove some attributes that can allows someone to identify the user.

15 What is the purpose of k-anonymity, l-diversity and t-closeness models? What is the relation between them?

The purpose of $k-anonymity$ is to "blur" data to separate it from the user it came from, from example by rounding values, or placing it in a scale; each record needs to be indistinguishable from at least $k-1$ other records.

$l-diversity$is the idea that each type of field need several data for each value. $k-anonymity$ prevent from identity disclosure, where $l-diversity$ prevent attribute disclosure (if all records with attribute A also have attribute B then finding someone with attribute A let us know they also have attribute B).

$t-closeness$ takes in account the original proportion of an attribute, where $l-diversity$ does not. For example, in cybersecurity school, there is a lot less women than men, therefore having a group from cybersecurity school with half women and half men would gives us information on this group (there is much more chance to find a women than in a cybersecurity school globally). The aim of $t-closeness$ is to reduce as much as possible those type of information gain, computing the distance between that sub-group and the global group. If the distance is too high (treshold to define), then the information gain is to high to satisfy the $t-closeness$.