A complete k-anonymity scenario

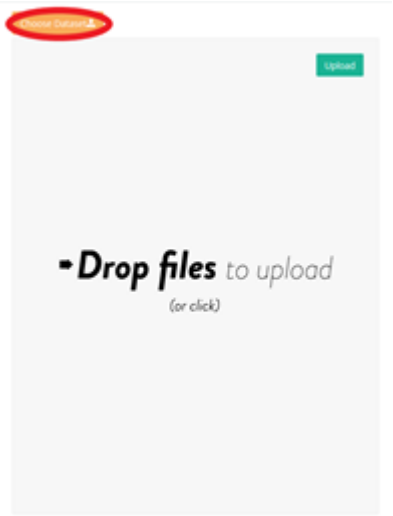
The aim of this tutorial is to provide an end-to-end demonstration of how to anonymize a dataset using an algorithm that provides k-anonymity.

Loading a dataset

The first step for anonymizing a dataset *D* is to load *D* in Amnesia. Note, that Amnesia takes as input the original non-anonymous dataset *D*. If Amnesia is not used locally, then *D* is transmitted over the network to the Amnesia server and it is susceptible to attacks from malicious 3rd parties.

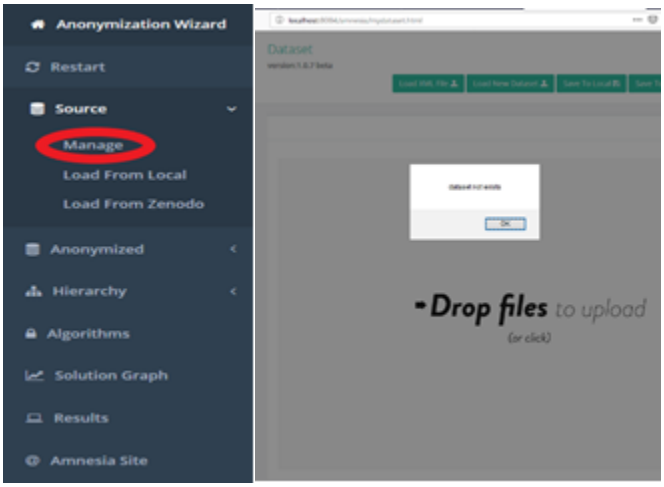
The dataset D can be uploaded in the following ways:

Option 1: Click the orange button and then choose D to upload.

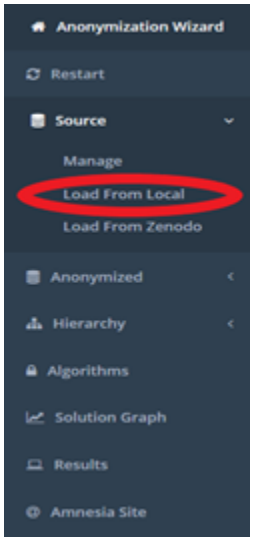


Option 2: Click in order to choose *D* or drop it and then click the “upload button”.

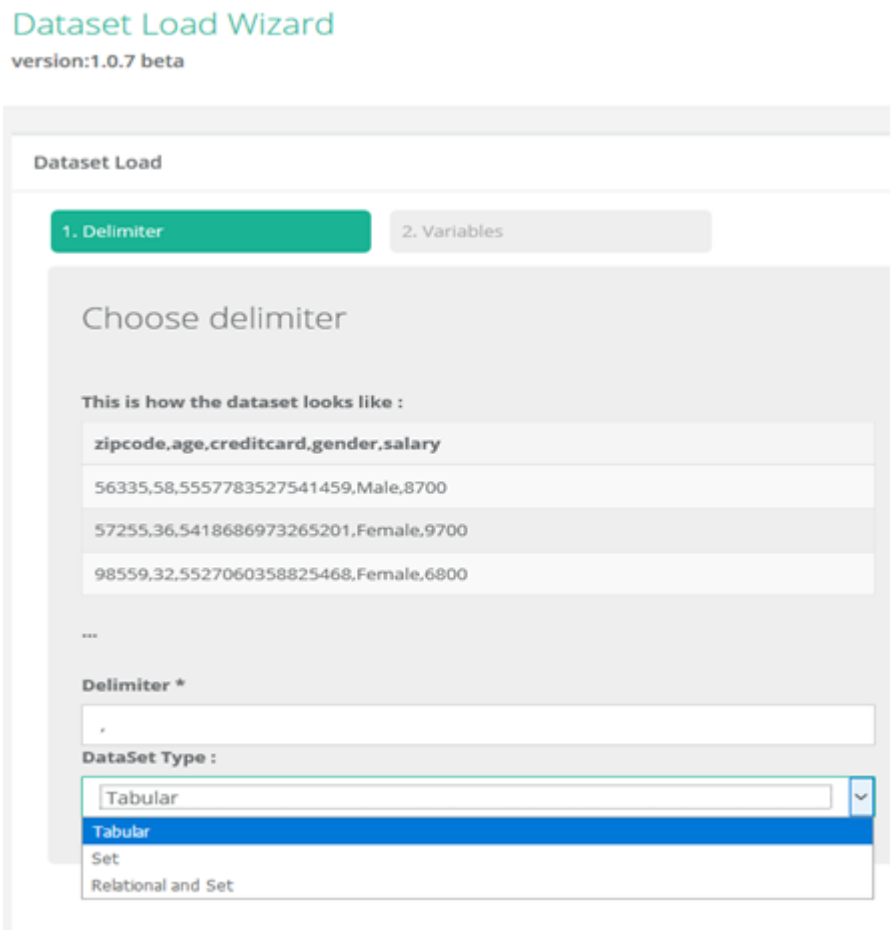
Option 3: Click the “source” option in the left menu and then “manage”, read the warning message after that click “load new dataset” and attach *D*.



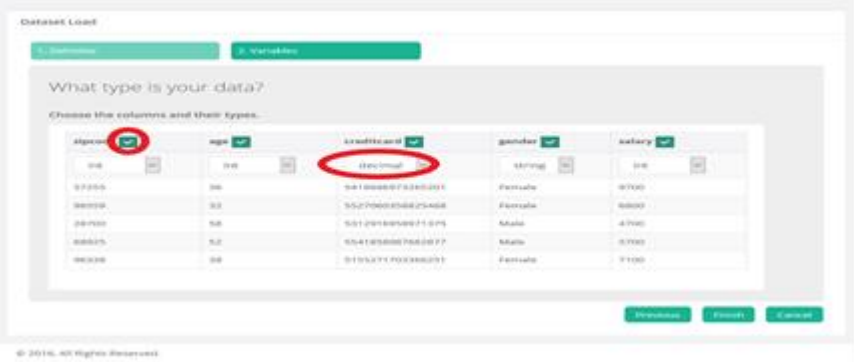
Option 4: Click the “source” option in the left menu and then “load from Local” in the end choose the dataset she/he wants to upload.



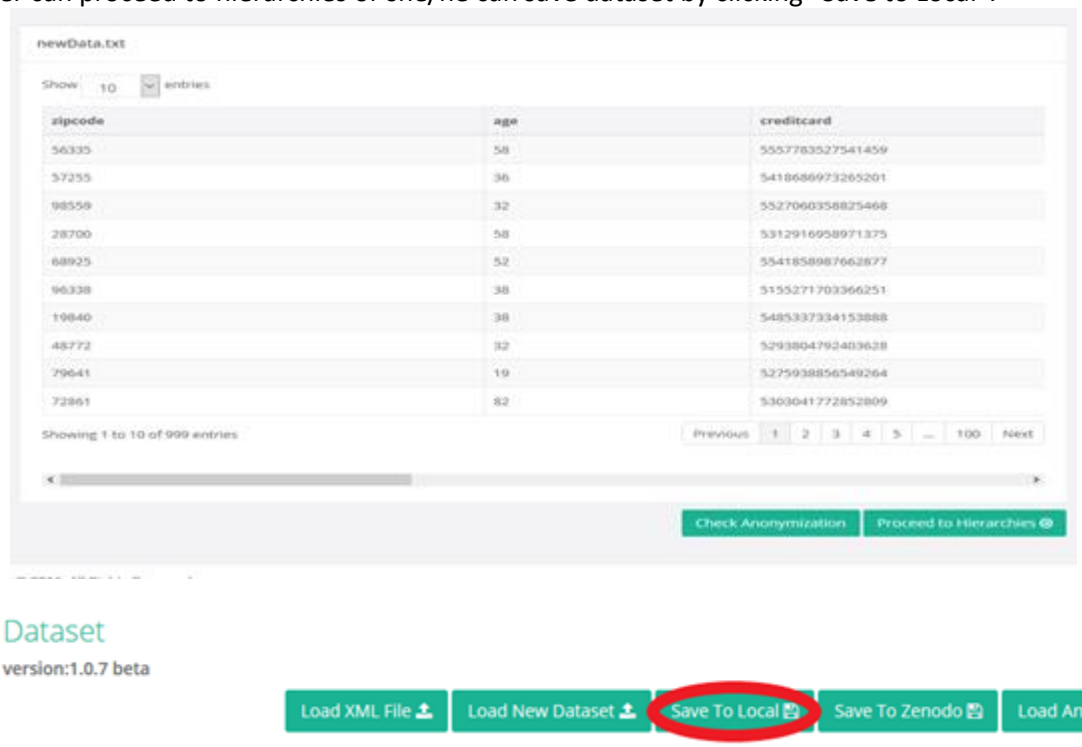
Amnesia accepts input text files where records span a single line, and distinct values are separated by a fixed delimiter. K-anonymity is applied only on tabular data, i.e., data that have a fixed number or values in each record. This means that D must have a fixed number of columns. Once D is uploaded to Amnesia, the user must guide the import wizard so that D is properly interpreted. The user must indicate the delimiter and the type of the dataset (sets and relational-set datasets can only be anonymized with k^m-anonymity). In this case the delimeter is the comma "," .



Amnesia presents all columns in the dataset and guesses their type. The user must verify the type of each column and mark the columns that will participate in the anonymization procedure.



When all choices are made the user must click the Finish button to load the dataset. Next, the user can proceed to hierarchies or she/he can save dataset by clicking “Save to Local”.

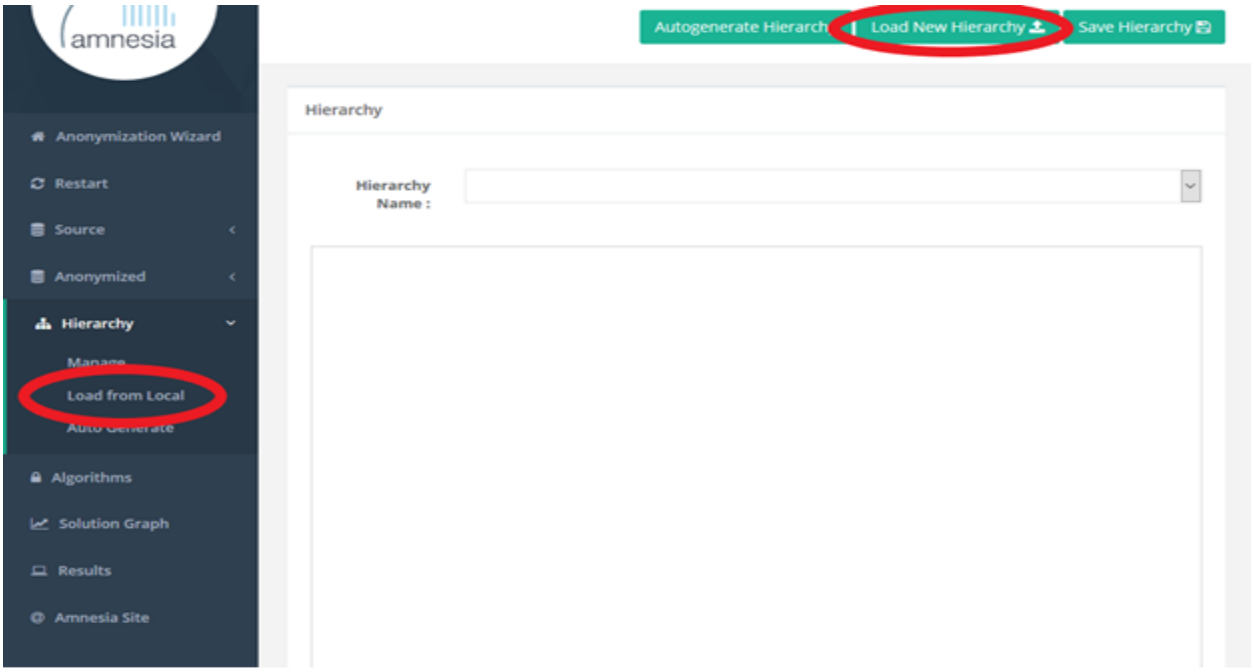


Introducing Generalization hierarchies

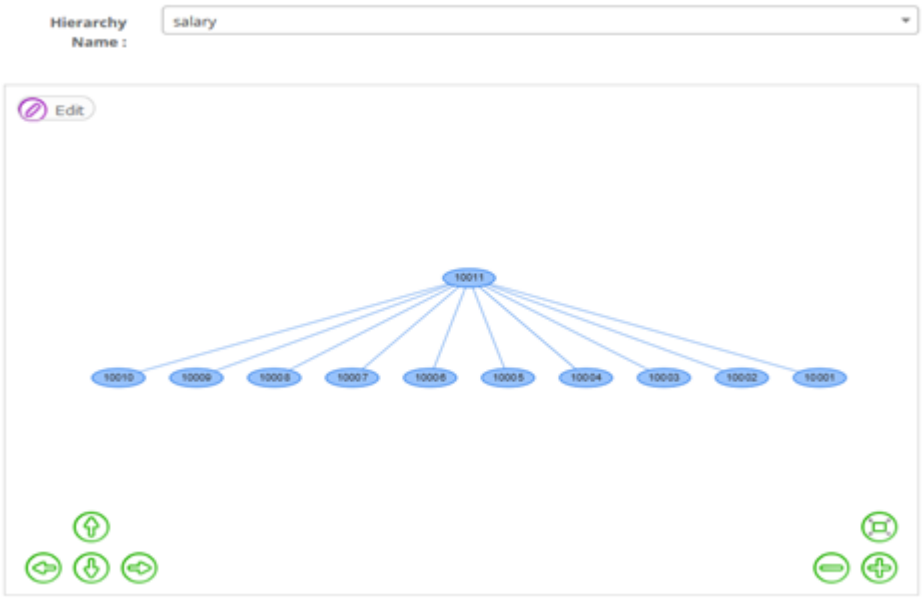
Amnesia’s k-anonymity algorithm uses generalization to create groups of identical data. Unique or rare (appearing less than k times in the dataset) values or value combinations are replaced by more generic values. For example, if there is a single person in a dataset that resides in Greece, then Greece (and the rest of EU countries) will be replaced by EU to create more than k identical values in the Country of Residence value. Generalization hierarchy that define how these replacements take place are provided by the user as input to the algorithm.

Amnesia assists the user to create hierarchies using existing datasets, more details on how to create and edit and existing hierarchy are provided in the XXXXXX Tutorial

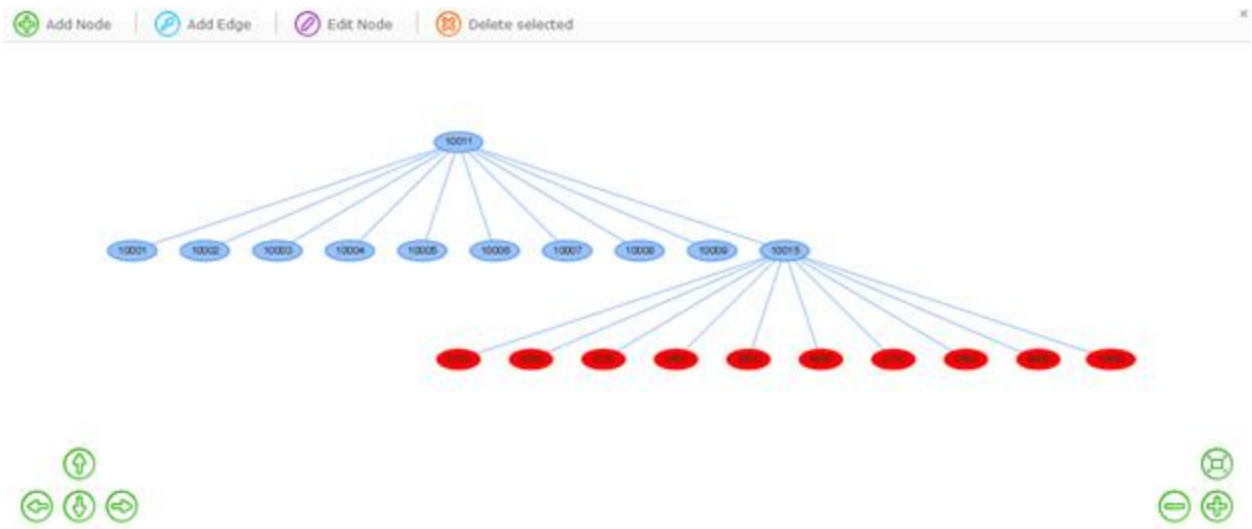
The user can upload a hierarchy in 2 different ways: a) by clicking “Load New Hierarchy” button or b) by clicking “Load from Local” in the left side menu.



Loaded hierarchies appear on the screen as graphs, where its parent node indicates how its children nodes are generalized.

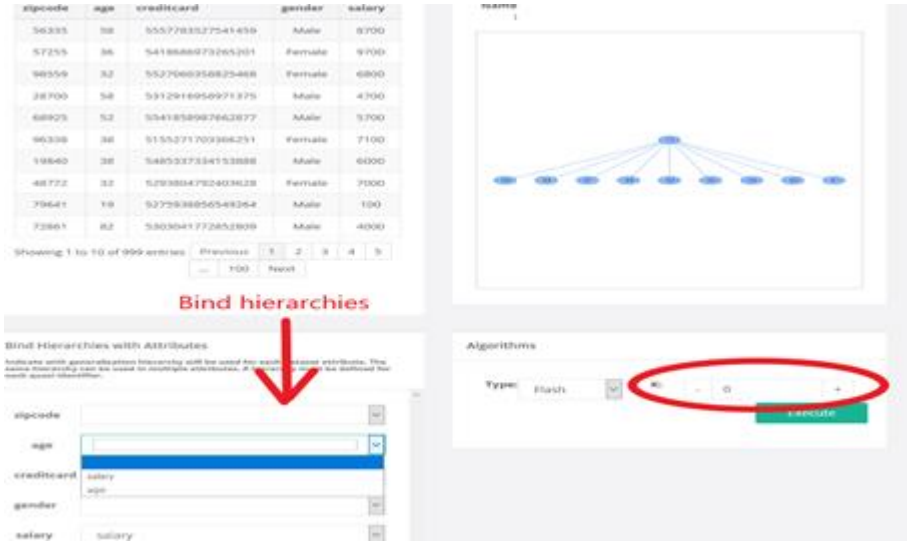


By double clicking in a node one can see its children.



Execution of Algorithms and Results

Once, the dataset and the hierarchies are loaded, the user can proceed with invoking the anonymization algorithm. The user only needs to: a) associate each attribute with the hierarchy that must be used on it and b) define the value of k. The execute button start the anonymization algorithm.



Solution space

The anonymization algorithm does not automatically choose a solution but offers all the possible solutions to the user. Solutions are presented graphically as a lattice, where each node represents a different solution. Blue nodes satisfy the k-anonymity requirements and red nodes do not. Each node is associated with a sequence of numbers. Each number shows the generalization level of each quasi identifier. The first number shows the generalization level of the first quasi identifiers, the second for the second quasi

identifier etc. More details are presented if the user hovers the pointer over a node. By clicking on a node a pop-up menu is activated that offers 2 options: a) preview a sample of the anonymized dataset or a) see statistics about the anonymized dataset corresponding to the selected solution.

.

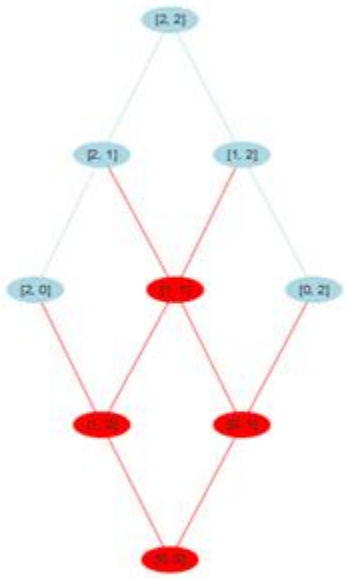
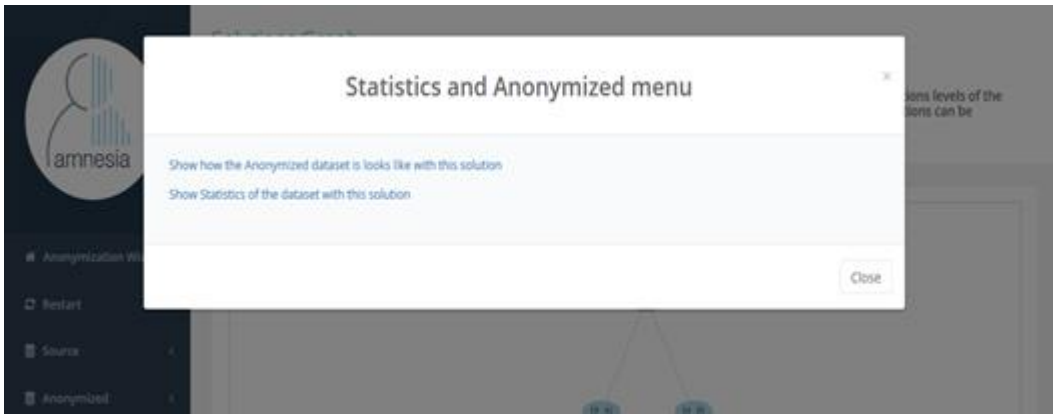
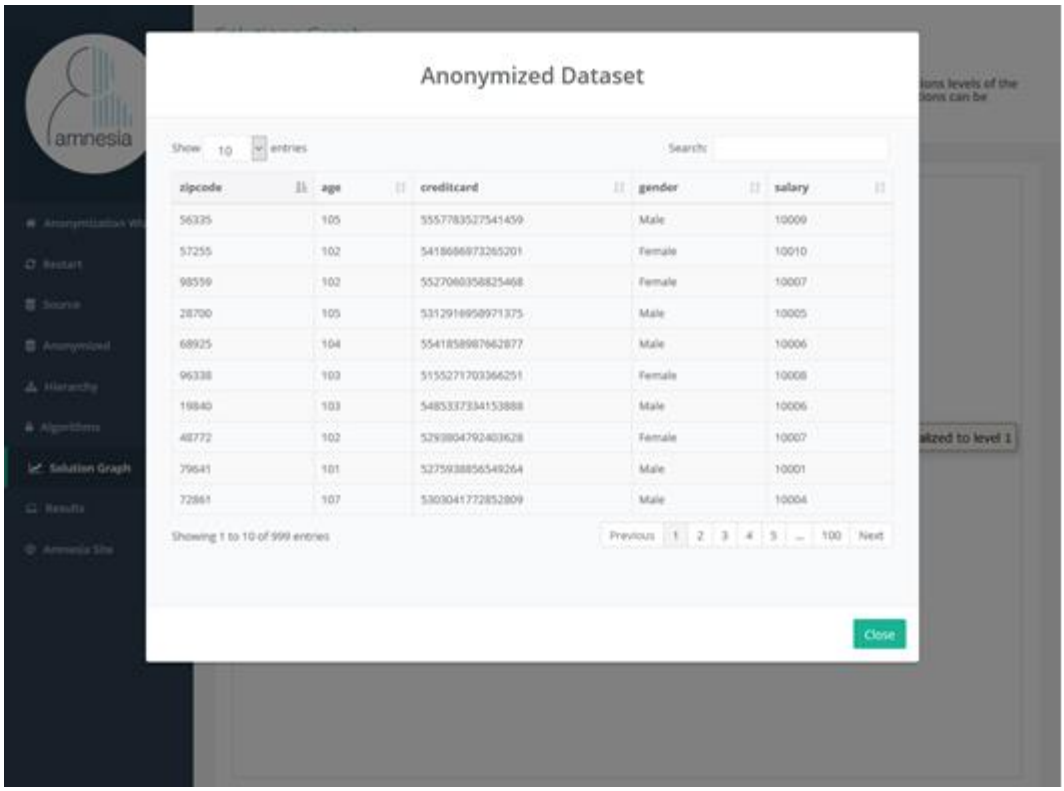
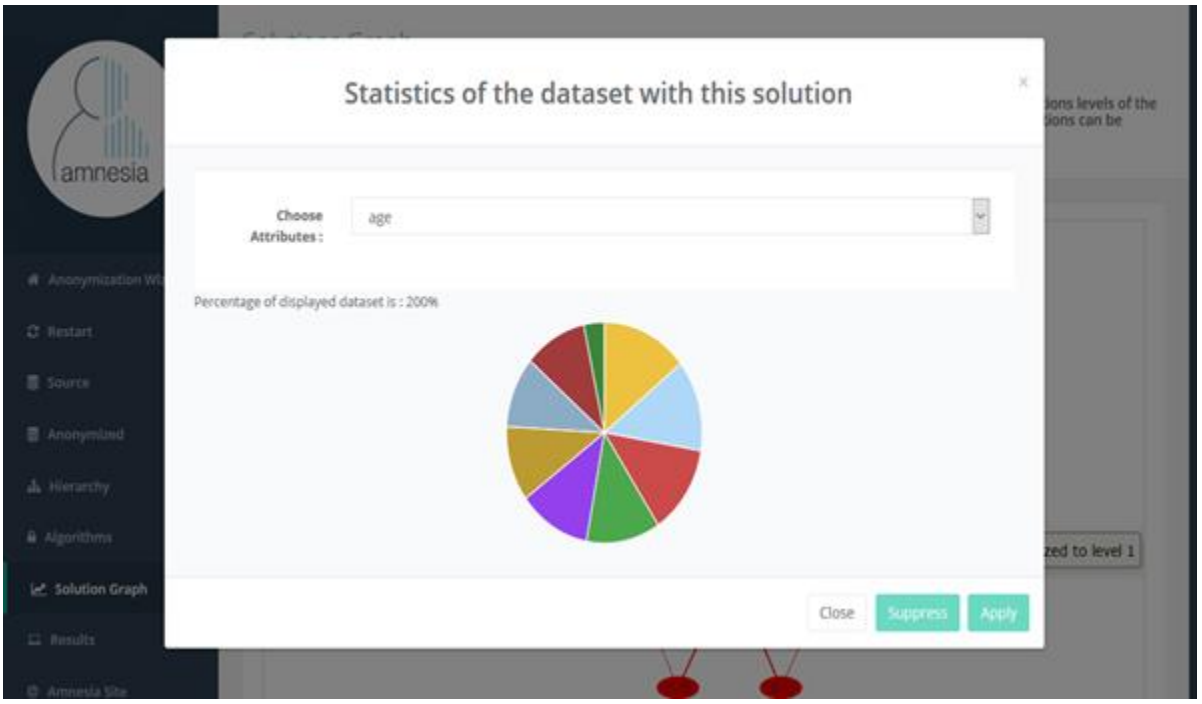


Figure 1 - The solution graph of the anonymization of two columns(salary, age) with k=4



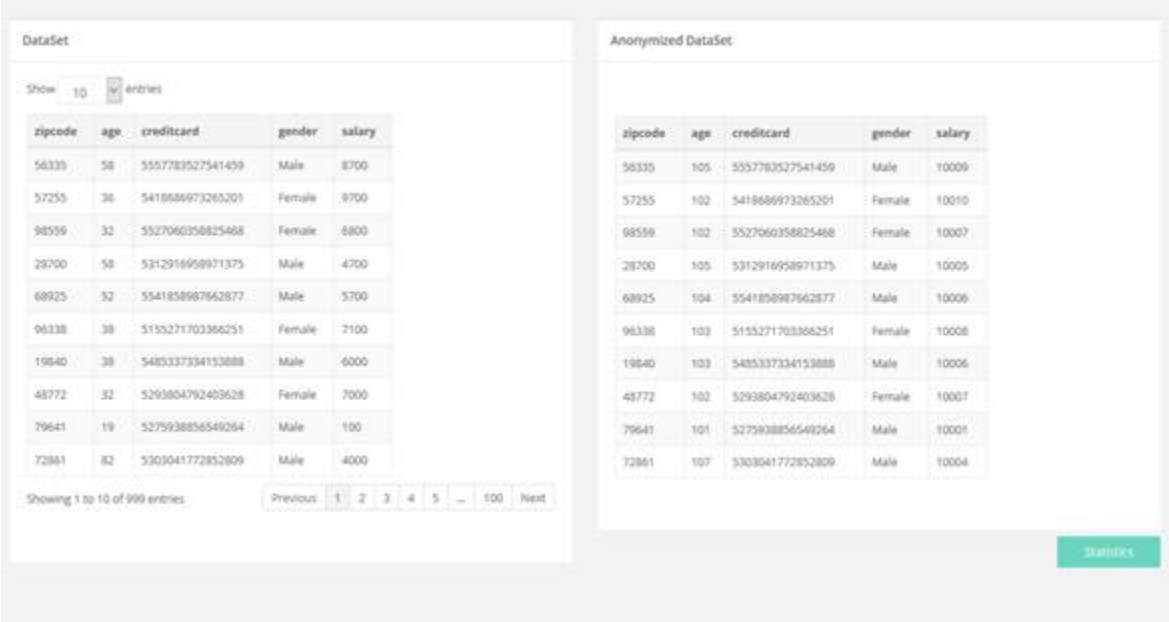




The idea behind depicting all possible solutions is to allow the user to choose the one that offers the best trade-off between privacy and the required utility. Whereas all blue nodes satisfy the privacy requirements, different nodes generalize different attributes to different levels.

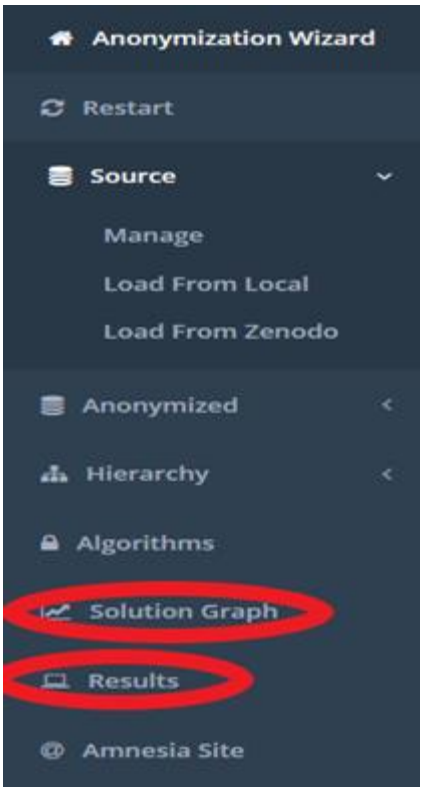
For example, one might generalize *salary* more that *age*, whereas another might do the opposite, and both provide the required privacy guaranty. A user can choose one depending on the intended use. Moreover, Amnesia offers one additional way to reduce information loss. Solutions that do not meet the privacy requirements with generalization, i.e., they appear as red nodes in the graph, might violate the requirement only because of a few records, e.g., 0.13% of the total records. The user can see the number of records that violate the privacy guaranty in the Statistics screen and then choose to suppress them (by clicking “Suppress” button), if she or he believes that more information is preserved by losing a few records in order to keep the generalization level low.

By double clicking a node the anonymization is applied and the user will be able to see the anonymized dataset side by side to the original dataset.



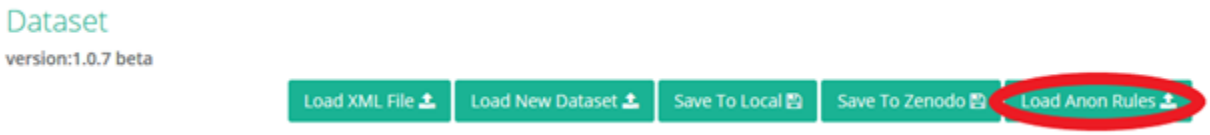
In this case you can save the anonymized dataset or additionally can save the anonymization rules. The anonymization rules are a representation of the selected solutions.

This screen can be reached directly by clicking the results option in the left side menu.



Anonymization rules

Stored anonymization rules can be imported and applied to other datasets that have similar data and the dataset will be directly anonymized. After loading the dataset the user has the option to load anonymization rules as it can be seen below.



At this point two scenarios are presented which handle set-value and object-relational datasets. In short, the procedure is simple, the user should load the appropriate dataset, load the hierarchies bind the hierarchies with the appropriate attributes, run the algorithm and finally the anonymized dataset is appeared on his/her screen next to the initial dataset.