

A complete k-anonymity scenario

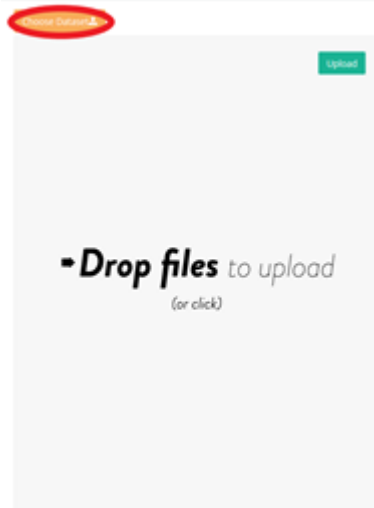
The aim of this tutorial is to provide an end-to-end demonstration of how to anonymize a dataset using an algorithm that provides k-anonymity.

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:

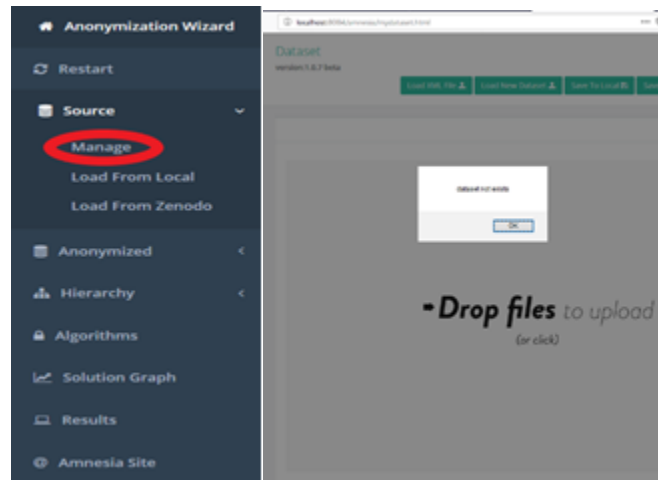
Option 1: Click the orange button and then choose D to upload.



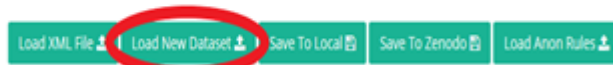
Option 2: Click in order to choose D or drop it and then click the “upload button”.



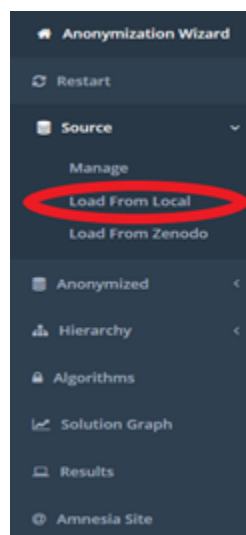
Option 3: Click the “source” option in the left menu and then “manage”, read the warning message after that click “load new dataset” and attach D.



Dataset
version:1.0.7 beta



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 delimiter is the comma “,” .

Dataset Load Wizard

version:1.0.7 beta

Dataset Load

1. Delimiter
2. Variables

Choose delimiter

This is how the dataset looks like :

zipcode,age,creditcard,gender,salary
56335,58,5557783527541459,Male,8700
57255,36,5418686973265201,Female,9700
98559,32,5527060358825468,Female,6800

...

Delimiter *

,

DataSet Type :

Tabular

Set

Relational and Set

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.

Dataset Load

1. Delimiter
2. Variables

What type is your data?

Choose the columns and their types.

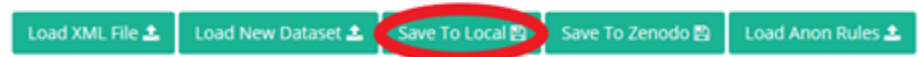
zipcode	age	creditcard	gender	salary
56335	58	5557783527541459	Male	8700
57255	36	5418686973265201	Female	9700
98559	32	5527060358825468	Female	6800

Previous
Next
Cancel

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”.

zipcode	age	creditcard
56335	58	5557783527541459
57255	36	5418686973265201
98559	32	5527060358825468
28700	58	5312916958971375
68925	52	5541858987662877
96338	38	5155271703366251
19840	38	5485337334153888
48772	32	5293804792483628
79641	19	5275938856549264
72861	82	5303041772852809

Dataset
version:1.0.7 beta

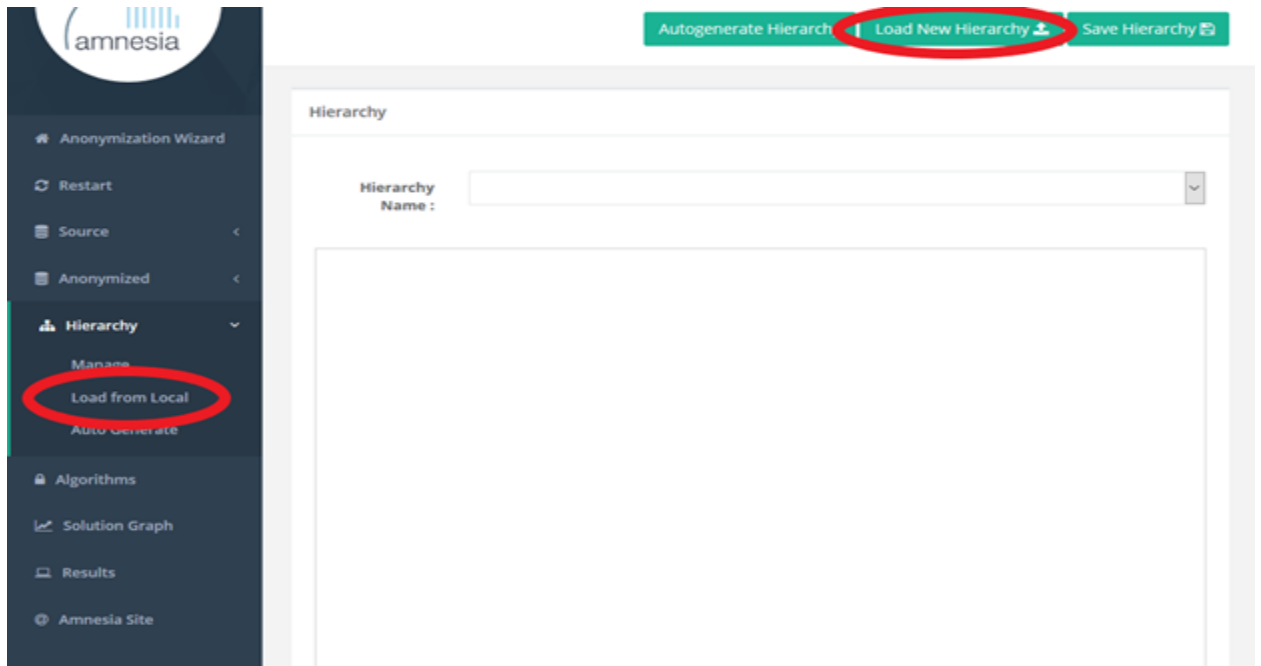


Introducing Generalization hierarchies

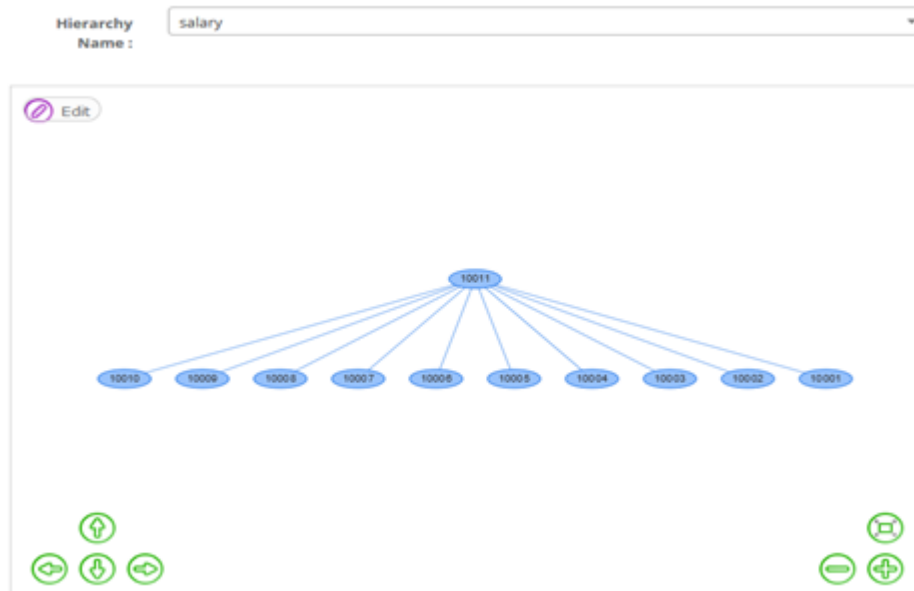
Amnesia’s k-anonymity algorithm uses generalization to create groups of identical data. Unique or rare (appearing less than k times in the dataset) values or value combinations are replaced by more generic values. For example, if there is a single person in a dataset that resides in Greece, then Greece (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](#)

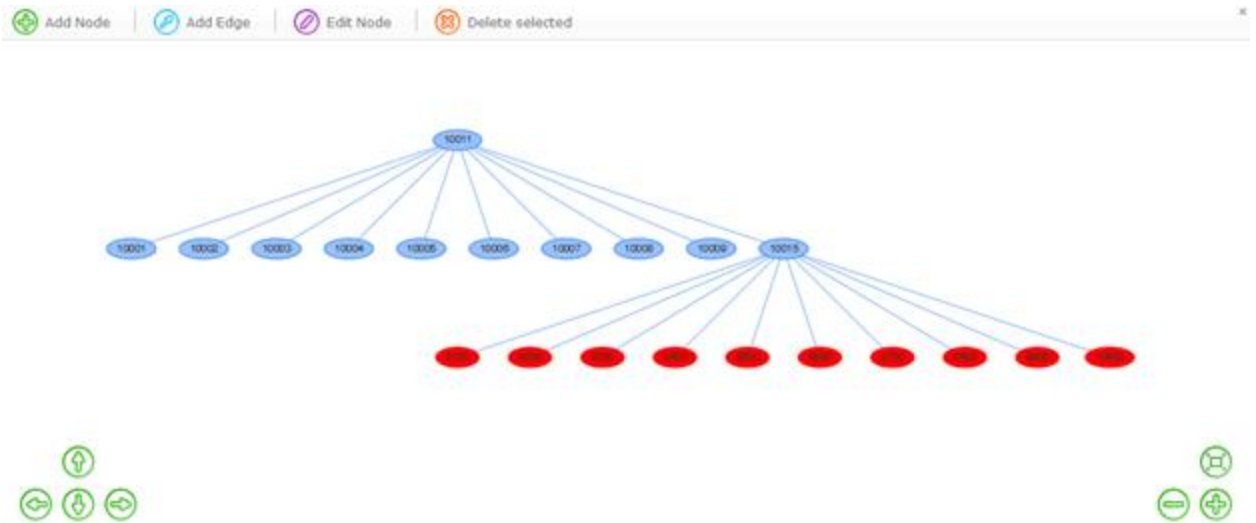
The user can upload a hierarchy in 2 different ways: a) by clicking “Load New Hierarchy” button or b) by clicking “Load from Local” in the left side menu.



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.

The screenshot shows a web application interface. On the left, there is a table with columns: zipcode, age, creditcard, gender, salary. The table contains 10 rows of data. Below the table, there is a 'Showing 1 to 10 of 999 entries' label and a 'Previous' button. In the center, there is a 'Bind hierarchies' section with a red arrow pointing to it. Below this section, there are dropdown menus for 'zipcode', 'age', 'creditcard', 'gender', and 'salary'. On the right, there is a 'name' section with a '1' label and a diagram of a hierarchy. Below this, there is an 'Algorithms' section with a 'Type: Hash' dropdown and a 'Execute' button circled in red.

zipcode	age	creditcard	gender	salary
56335	58	555728327541459	Male	9700
57255	36	5418888973265201	Female	9700
98559	32	5527068358825488	Female	6800
28700	58	5312916950971975	Male	4300
68925	52	5541858987662877	Male	5700
96338	38	5155271703388251	Female	7100
18840	38	5485337534153888	Male	6000
48722	32	5293884792403628	Female	7000
79641	18	52758388950548354	Male	100
72861	82	5303041772852809	Male	4000

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.

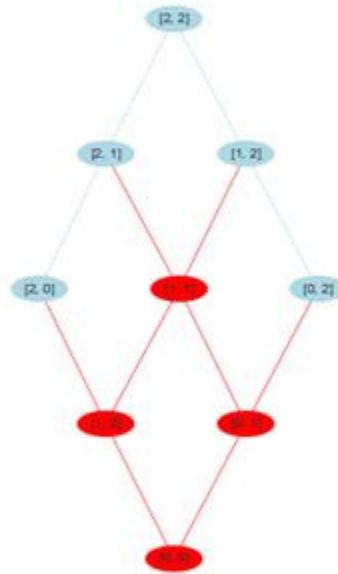


Figure 1 - The solution graph of the anonymization of two columns(salary, age) with $k=4$



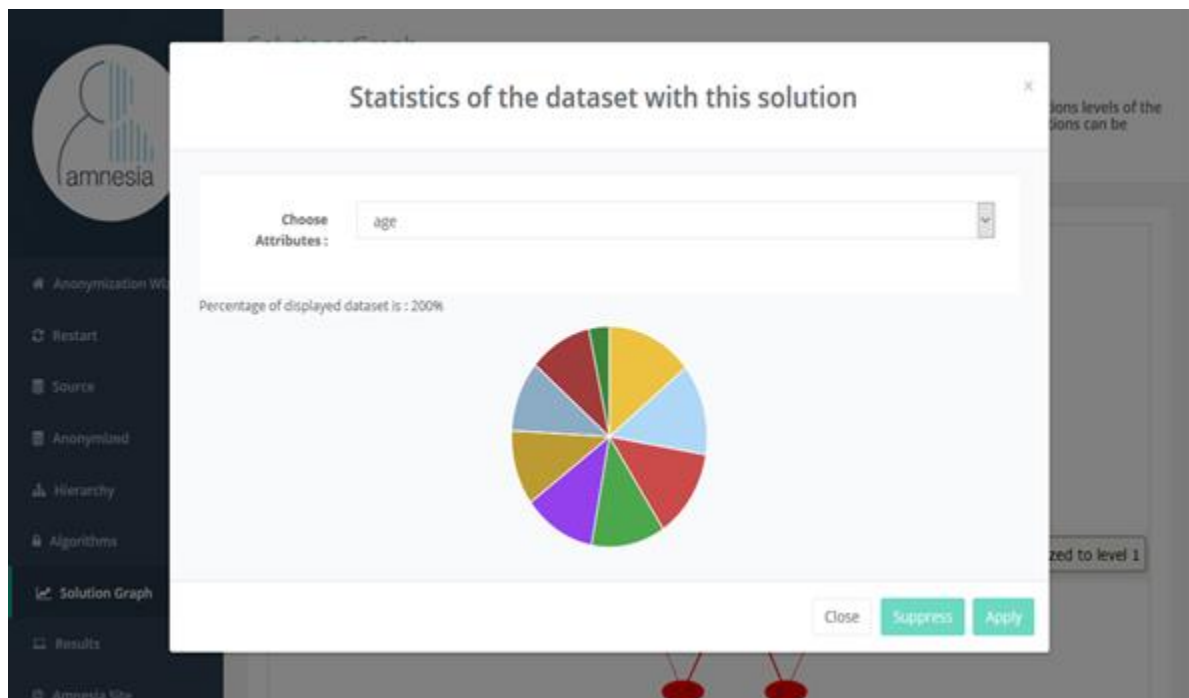
Anonymized Dataset

Show 10 entries Search:

zipcode	age	creditcard	gender	salary
56325	105	5557783527541459	Male	10009
57255	102	5418666873265201	Female	10010
98599	102	5527060358825468	Female	10007
28700	105	5312918958971375	Male	10005
68925	104	5541858987662877	Male	10006
96338	103	5155271703366251	Female	10008
19840	103	5485337334153888	Male	10006
48772	102	5299804792403628	Female	10007
79641	101	5275938836548264	Male	10001
72861	107	5303041772852809	Male	10004

Showing 1 to 10 of 999 entries Previous 1 2 3 4 5 ... 100 Next

[Close](#)



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 than *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.

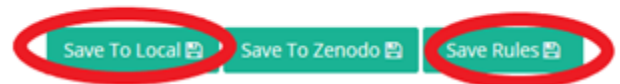
zipcode	age	creditcard	gender	salary
56335	58	5557783527541459	Male	8700
57255	36	5418686973265201	Female	9700
98559	32	5527060358825468	Female	6800
28700	58	5312916958971375	Male	4700
68925	52	5541858987662877	Male	5700
96338	38	5155271703366251	Female	2100
19840	38	5485337334153888	Male	6000
48772	32	5293804792403628	Female	7000
79641	19	5275938856549264	Male	100
72861	82	5303041772852809	Male	4000

zipcode	age	creditcard	gender	salary
56335	105	5557783527541459	Male	10009
57255	102	5418686973265201	Female	10010
98559	102	5527060358825468	Female	10007
28700	105	5312916958971375	Male	10005
68925	104	5541858987662877	Male	10006
96338	103	5155271703366251	Female	10008
19840	103	5485337334153888	Male	10006
48772	102	5293804792403628	Female	10007
79641	101	5275938856549264	Male	10001
72861	107	5303041772852809	Male	10004

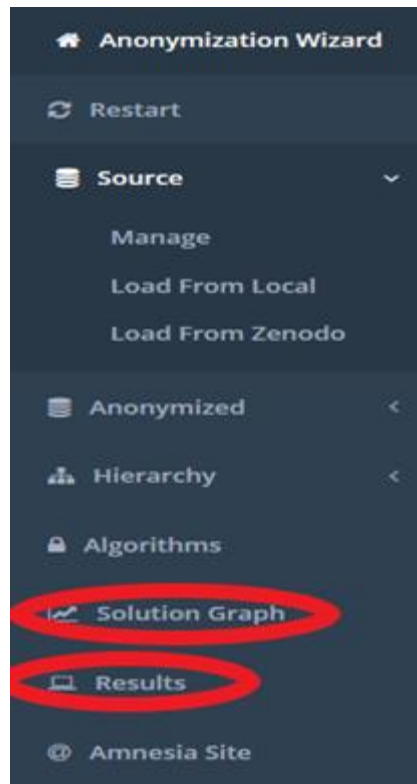
In this case you can save the anonymized dataset or additionally can save the anonymization rules. The anonymization rules are a representation of the selected solutions.

Results

version:1.0.7 beta



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

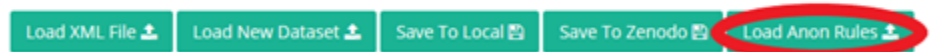


Anonymization rules

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.

Dataset

version:1.0.7 beta



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.