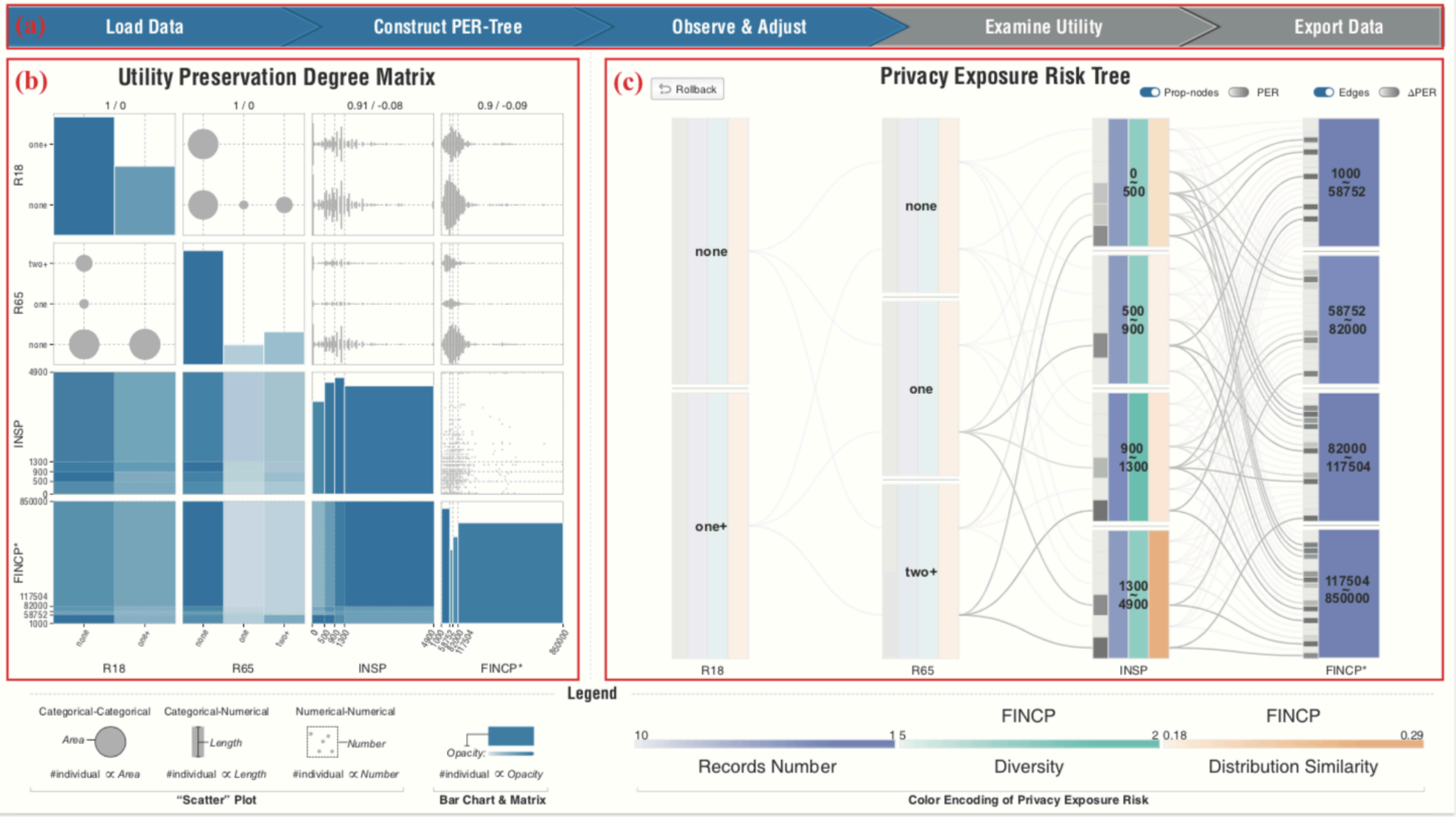
A Utility-Aware Visual Approach for Anonymizing Multi-Attribute Tabular Data

Xumeng Wang, Jia-Kai Chou, Wei Chen, Huihua Guan, Wenlong Chen, Tiny Lao, Kwan-Liu Ma

Shanice Clarke

Problem

- Making datasets available to others is beneficial for research.
- Attributes in a dataset can be used directly to specify an individual's identity.
- Before data can be shared it should be sanitized so sensitive information is not exposed.
- When data is removed or hidden it becomes less useful for analysis.
- Previous studies don't provide feedback of how much utility is reduced while handling privacy risks.



- Visual interface that helps users identify privacy exposure risk

- Allows users to see how the utility of the data is affected as they handle privacy issues

How to measure privacy?

- **Data is grouped into equivalence classes.**
- **k-anonymity**
  - There should be at least k records in a class with the same quasi-identifier.
  - A class is considered privacy-exposing if the number of data records, n, is smaller than k.
  - Degree of privacy exposure: k-n
  - Does not take the diversity of a sensitive attribute into account.
  - Information of individuals can still be revealed if they have the same quasi-identifier information and the same values for sensitive attributes.
- **l-diversity**
  - Data records in the same equivalence class must have l different values in the sensitive attribute.
- **t-closeness**
  - The distribution of sensitive values in each equivalence class must be smaller than a threshold, t.

How to measure utility?

- How to maintain and measure the utility of data has been a widely studied problem.
- Common approach:
  - Calculate the sum or average of the intervals size in the equivalence classes. Use this loss of information as the change in utility.

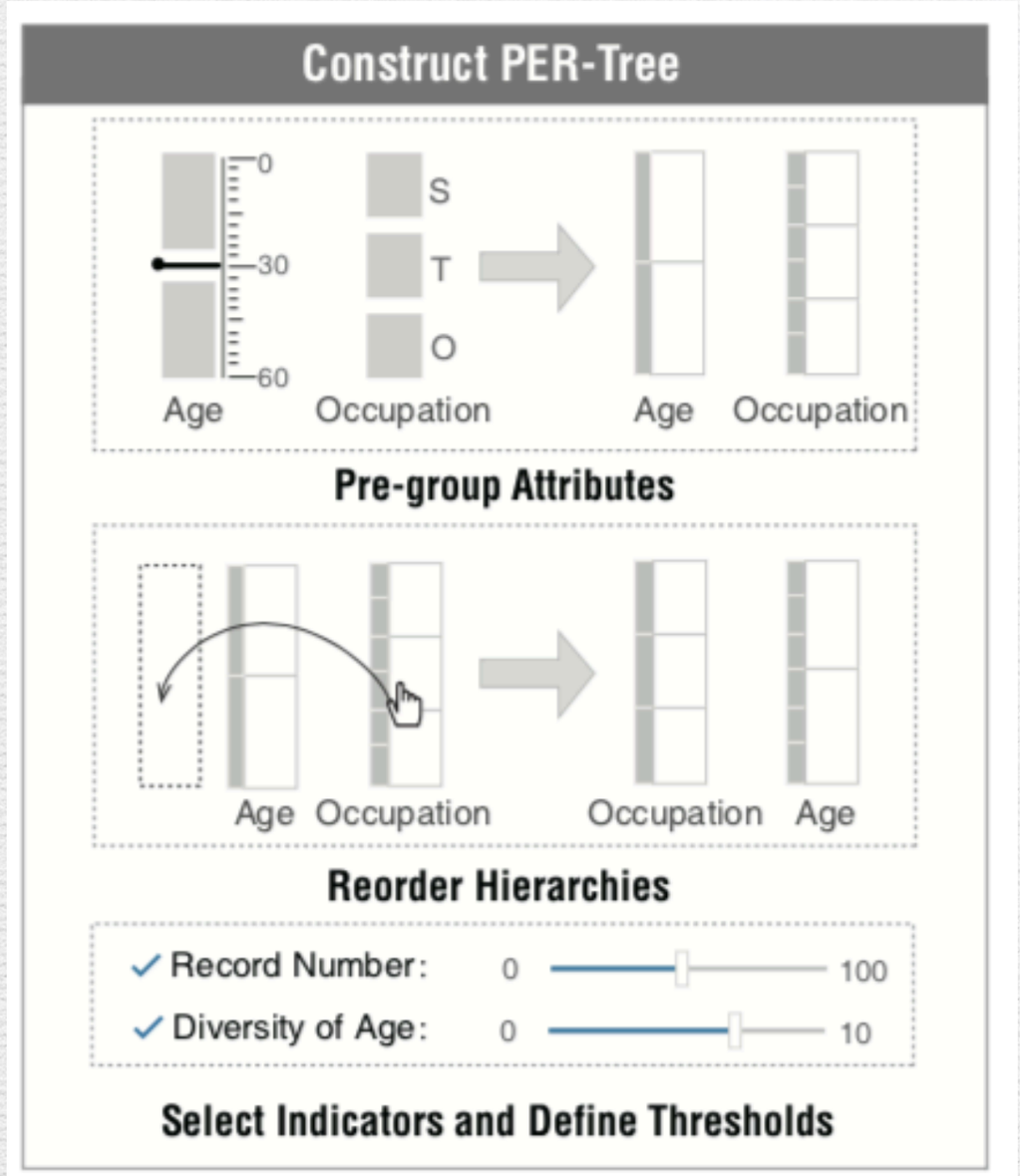
1. Load Data

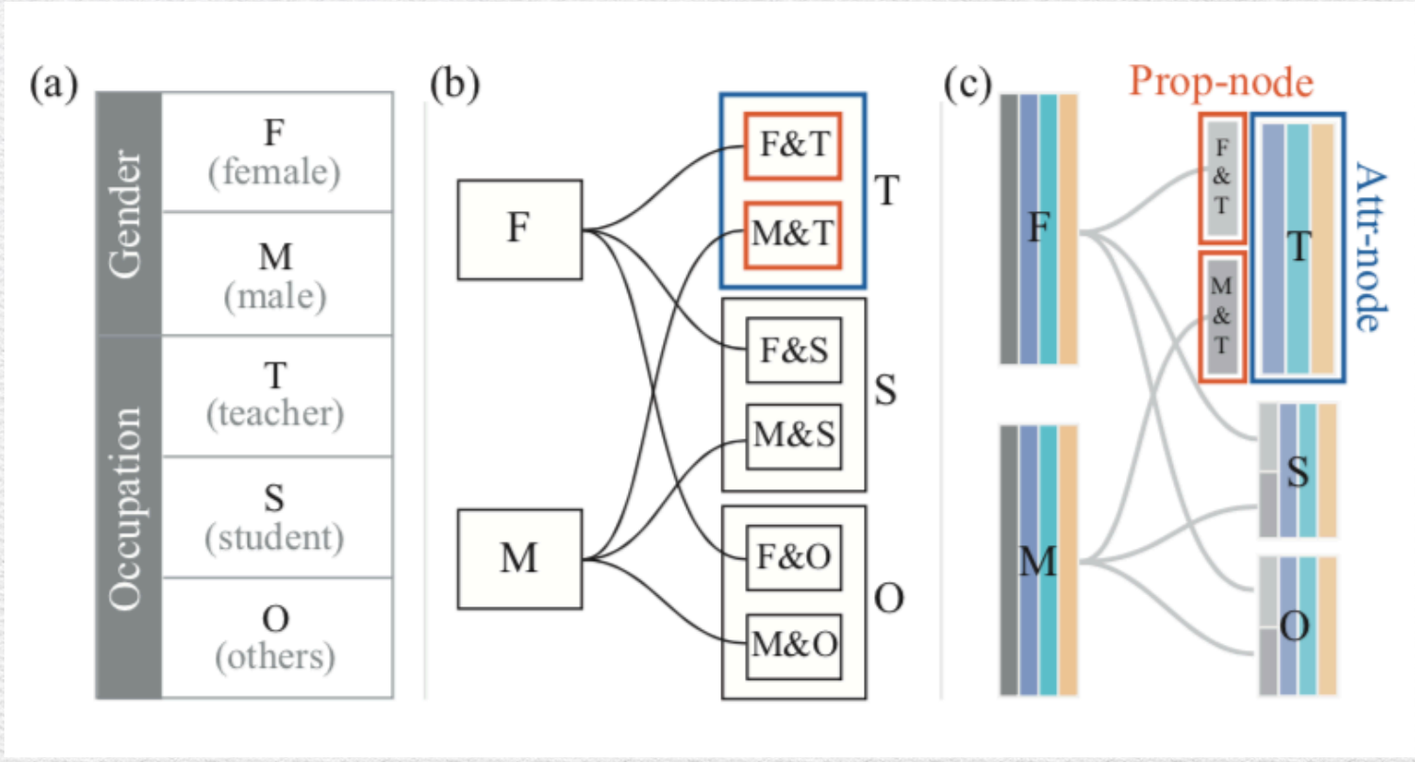
Load Data

Attr	Necessary	Sensitive	Description
Age	✓	🔓	between 0~100
Expense	✓	🔒	per month
Gender	✓	🔓	male/female
Name	+	🔓	first name, last name
Occupation	✓	🔓	teacher/student/other

Users select attributes of interest and indicate if they are require privacy protection.

2. Construct Privacy Exposure Risk Tree



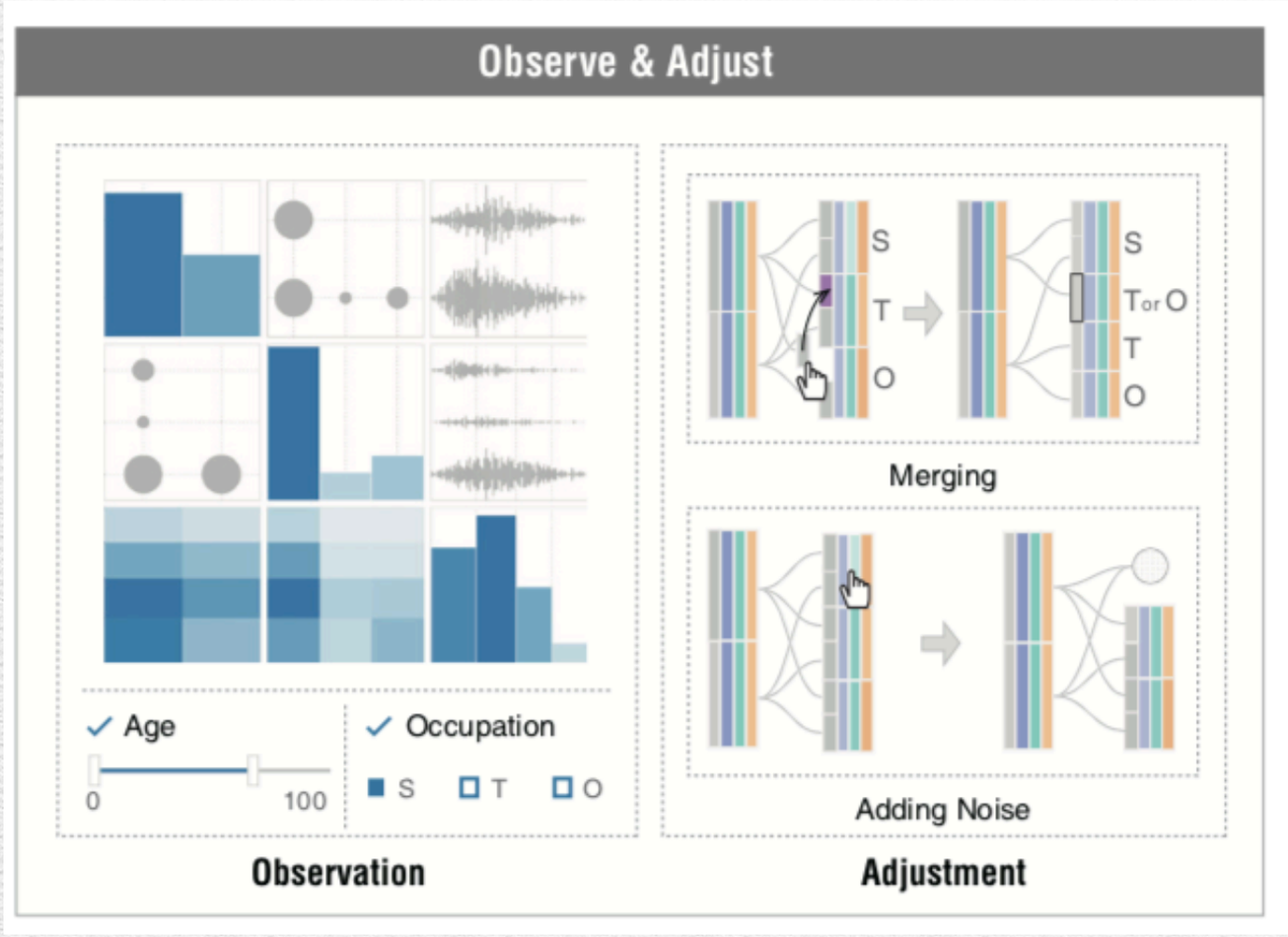


Attribute Node:
blue: k-anonymity
green: l-diversity
orange: t-closeness

opacity: maximum degree of privacy exposure

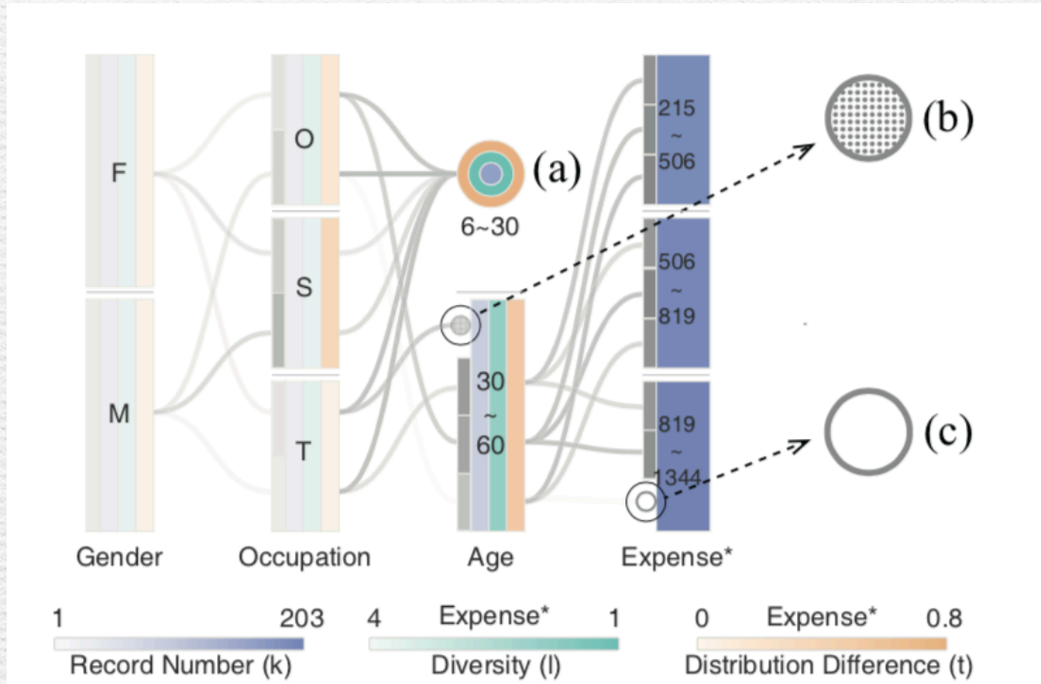
Prop Node:
opacity: total amount of privacy risk propagated from parent

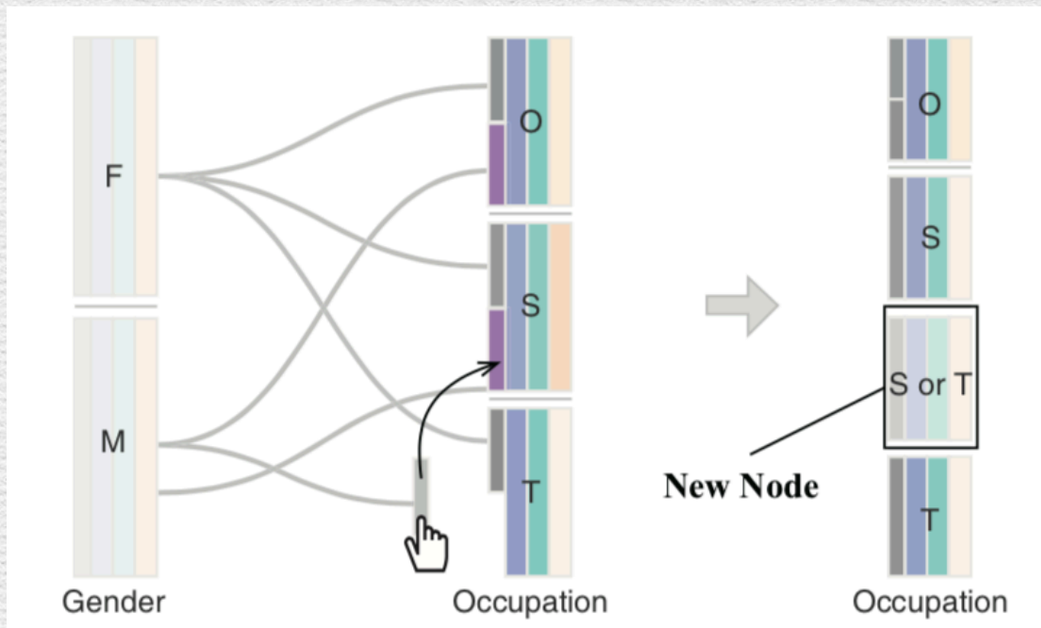
3. Observe and Adjust



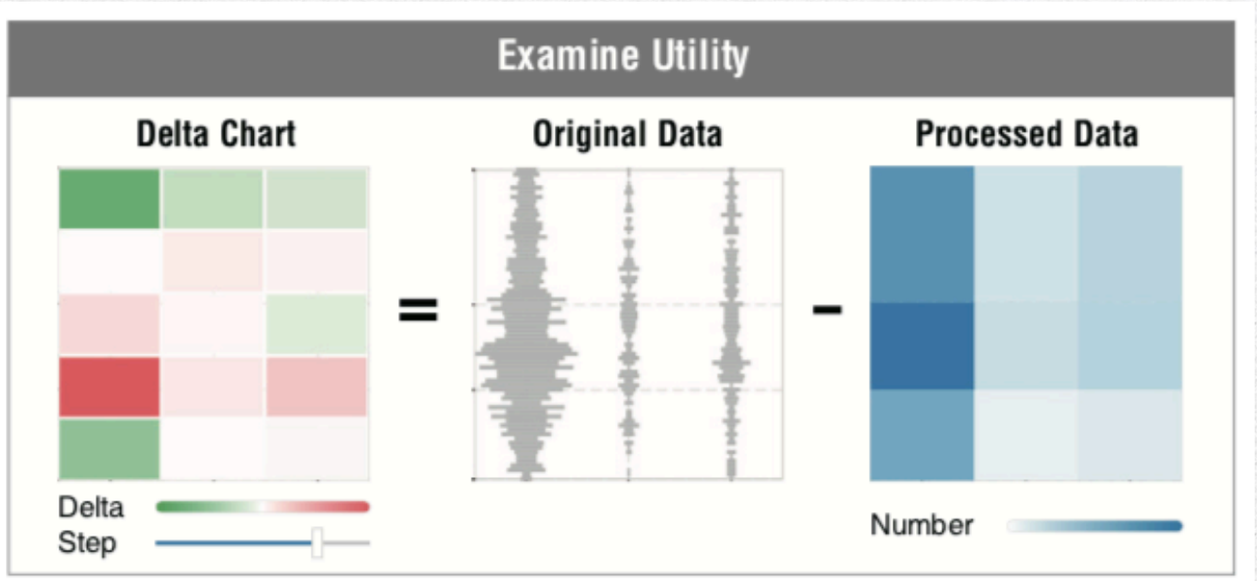
Utility Preservation Degree Matrix shows the distributions of attributes.

Nodes can be merged or collapsed.

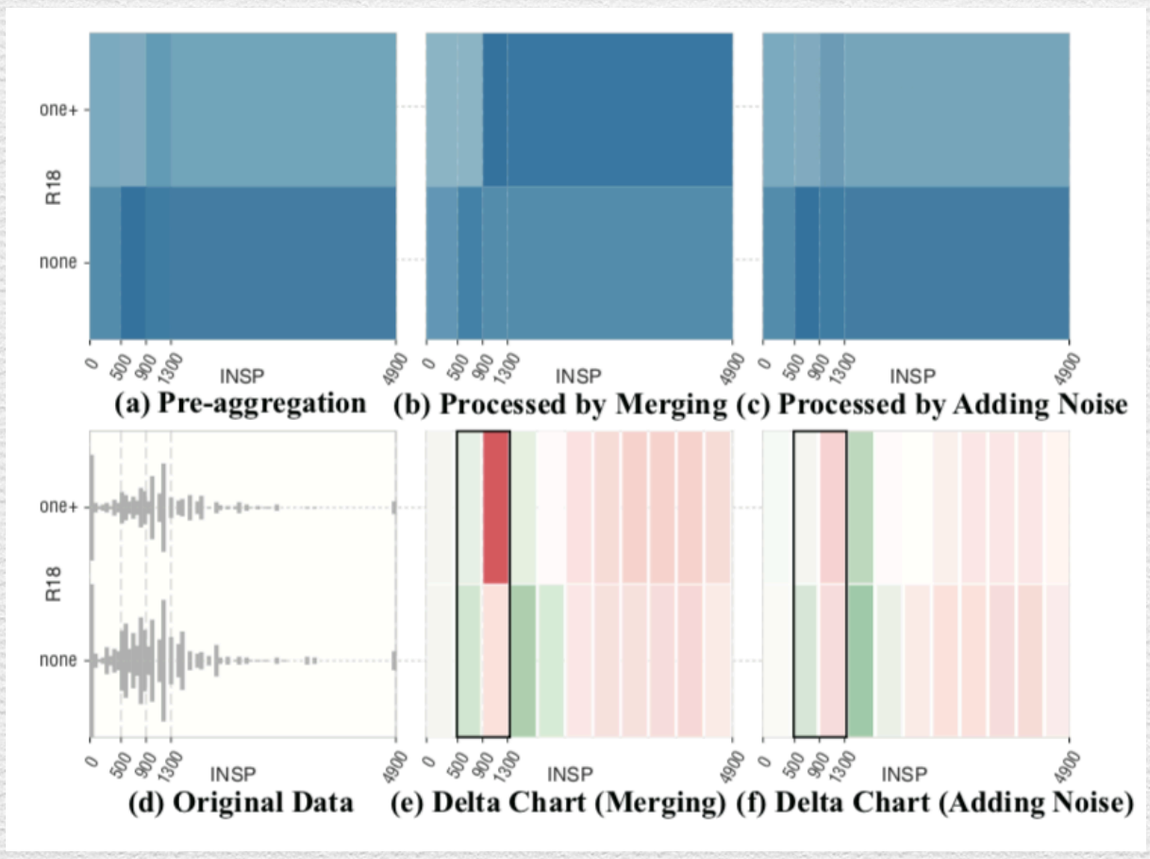




4. Examine Utility

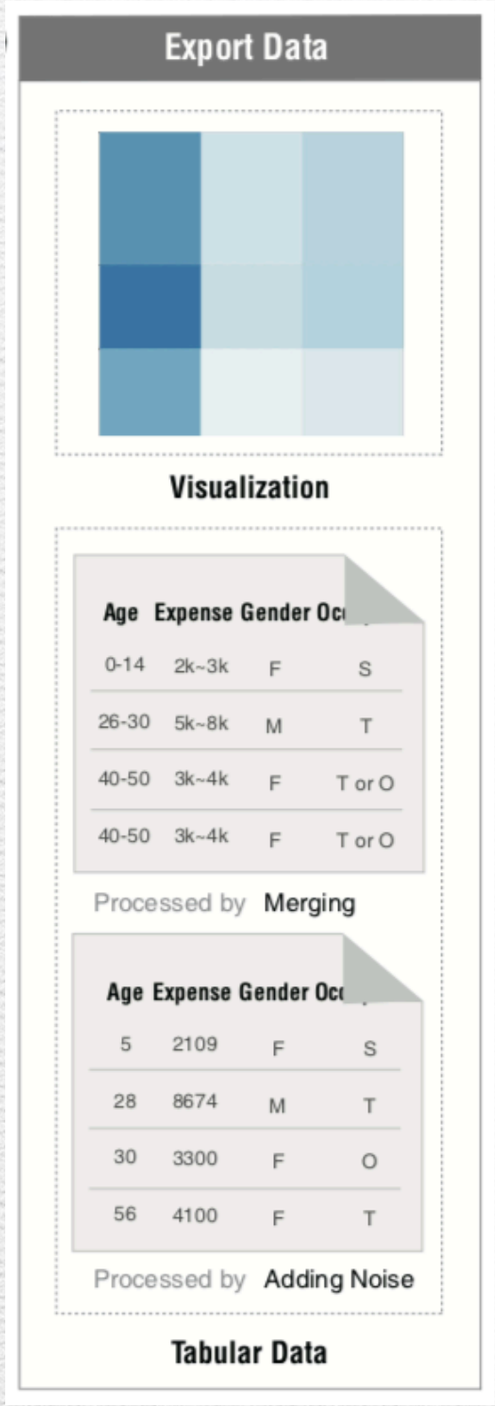


Users can examine the differences between the data before and after manipulations.



Delta charts shows the utility change from aggregation.

5. Export Data



Evaluation

- The system was evaluated by three experts who could be potential users. They liked:
  - the use of multiple measures of privacy
  - real-time feedback on changes in data utility

Improvements and Future Work

- The visual approach requires more time and engagement from the user than a data centric method.
  - A recommendation mechanism could be added to help users perform tasks quicker.
- The PER-tree suffers when there is a large number of attributes. Important privacy exposures could be missed.
  - Further interactions could be added to manipulate the display.
  - Provide alternate views to present the information in different ways.