A Utility-aware Visual Approach for Anonymizing Multi-attribute Tabular Data

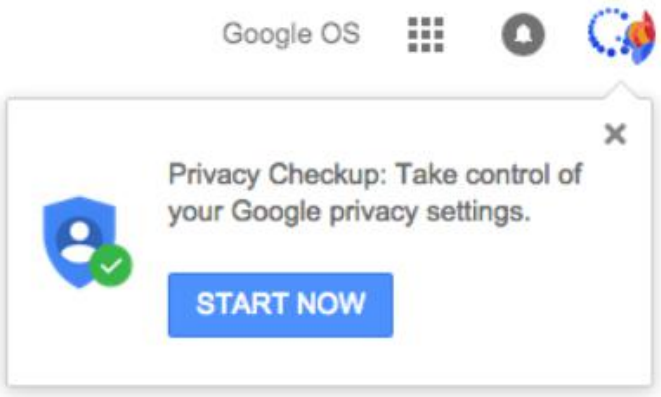
用于将多属性表格数据匿名化的实用性感知可视方法

王叙萌[1]；周家恺[2]，陈为[1]，关会华[1]，陈文龙[1]，劳天溢[1]，马匡六[2]

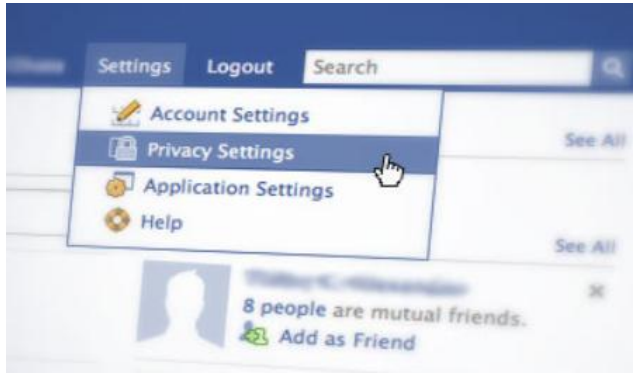
1: 浙江大学计算机辅助设计与图形学国家重点实验室

2: University of California, Davis

隐私保护的需求



Google



Facebook

微信

实用性维护的需求

实用性：传递信息的能力

如何平衡隐私和实用性？

可视化的作用

- 和自动方法相比，可视化更擅长于：
  - 解释过程
  - 整合专家知识
  - 根据问题定制解决方法

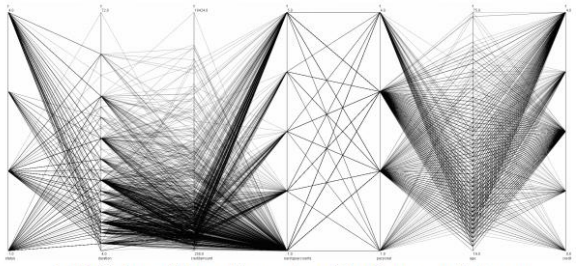
相关工作：隐私保护模型

- 语义匿名模型
  - K-anonymity [IJUFKS 2002]
  - L-diversity [IEEE ICDE 2006]
  - T-closeness [IEEE ICDE 2007]
- 差分隐私模型 [IEEE FOCS 2007]

相关工作： 隐私 vs 实用性

- 实用性定义
  - 结果的精确度 [ACM SIGKDD 2008, VLDB 2007]
  - 结果和原始数据之间的距离 [FAST 2011]
- 寻求平衡
  - 主要靠优化方法 [SIAM 2012]

相关工作：隐私相关可视化



(a) Default parallel coordinates view of the German credit dataset

(b) Clustered view with $k = 3$

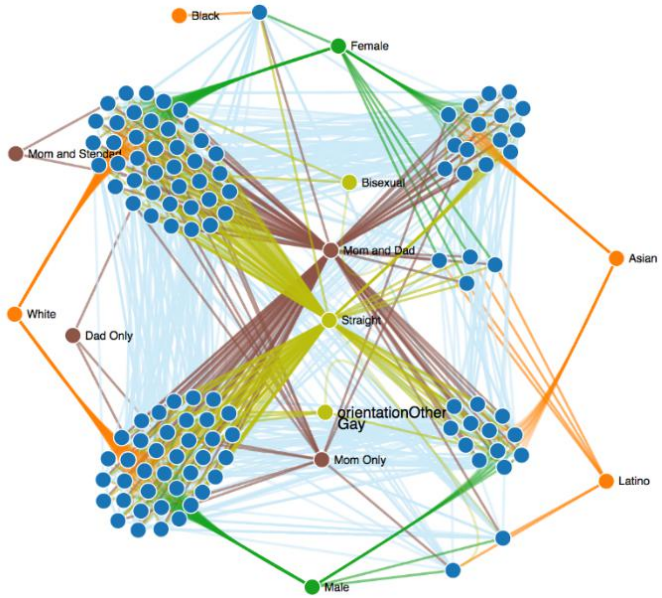
(c) Clustered View with $k = 6$

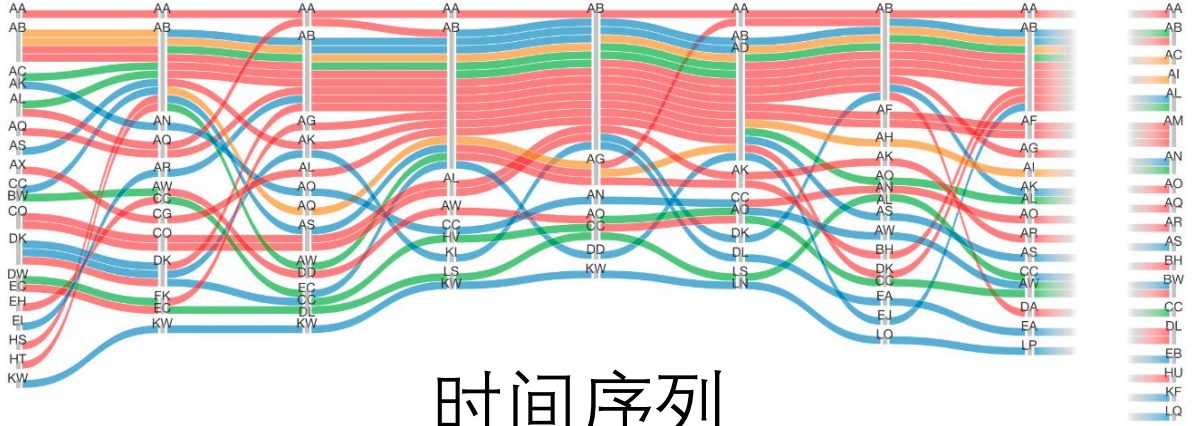
Screen-space Sanitization

[IEEE TVCG 2011]

图数据

[IEEE PVIS 2017]





时间序列

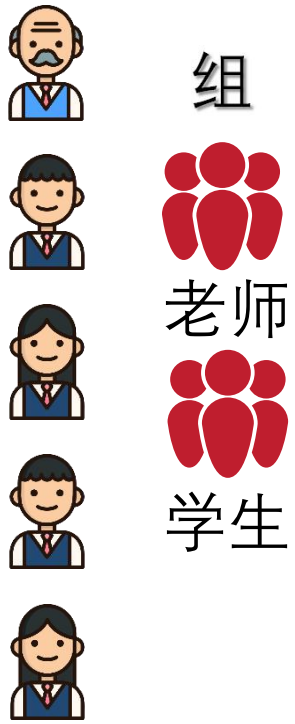
[ACM SIGGRAPH ASIA 2016]

VIS 2017

背景：语义匿名模型

必要属性

姓名	职业	收入	敏感属性
李磊	老师	26	
韩梅	学生	18	
张易	学生	21	
王帆	学生	25	
孙柏	学生	17	



职业	收入	总数
老师	20-30	1
学生	10-20	2
	20-30	2

K-anonymity

每组中至少有 k 个个体。

考虑 2-anonymity?

职业	收入	总数	
老师	20-30	1	不满足
学生	10-20	2	满足
	20-30	2	

L-diversity

每组至少有 l 个不同的属性值。

考虑 2-diversity?

职业	收入	总数	
老师	20-30	4	不满足
学生	10-20	2	满足
	20-30	2	

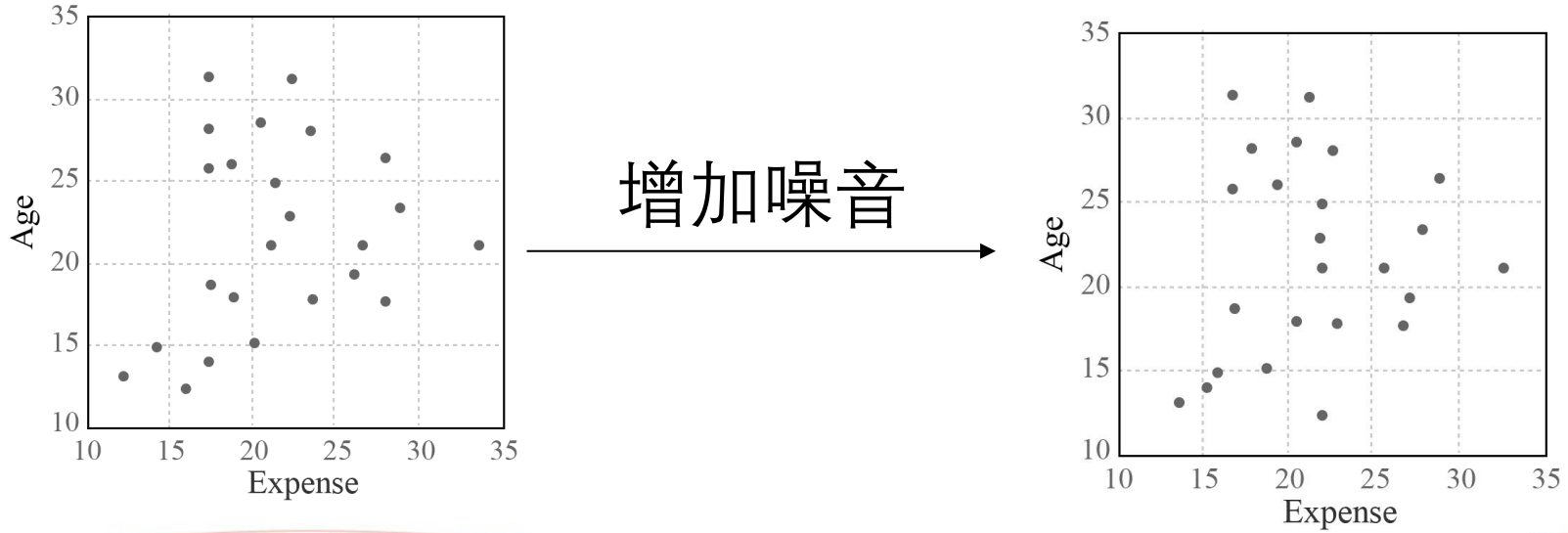
T-closeness

每组和整个数据集的敏感属性分布差异不能超过 t。

职业	收入	总数
老师	20-30	9
	30-40	1
学生	10-20	2
	20-30	2

背景： 差分隐私模型

- 给属性值添加随机噪音
- 保持整体分布近似不变



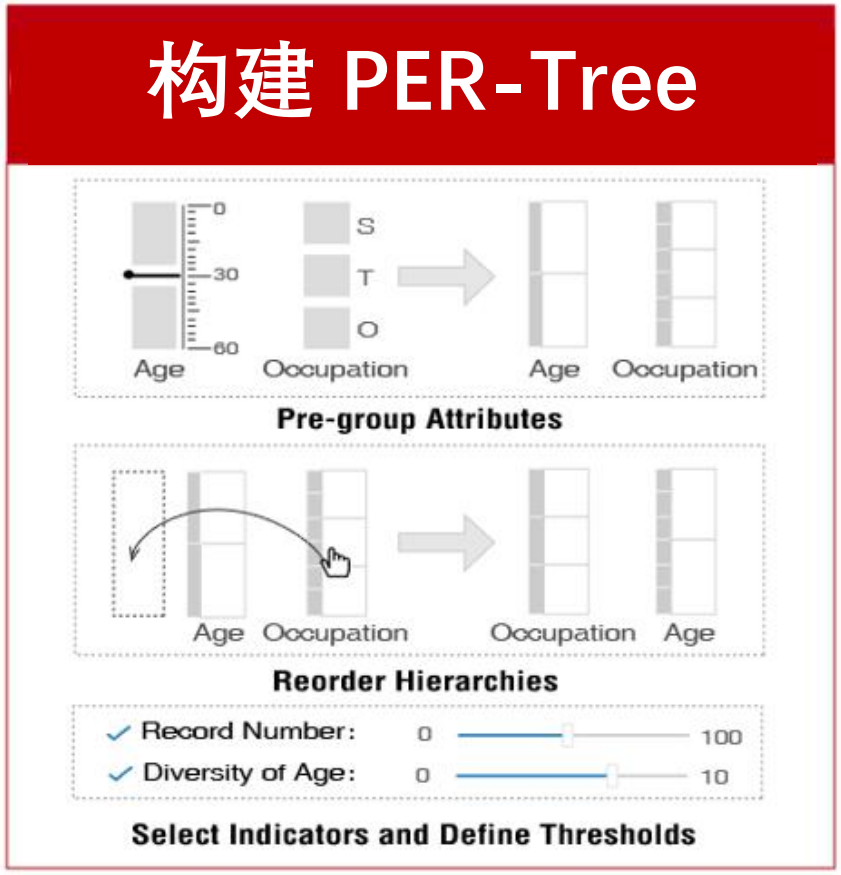
目标

- 数据
  - 多属性表格数据 —— 如何**表达**表格数据？
- 目标
  - 兼顾实用性的隐私保护
    - 如何**高亮**隐私问题？
    - 如何**解决**隐私问题？
    - 如何**评估**实用性？

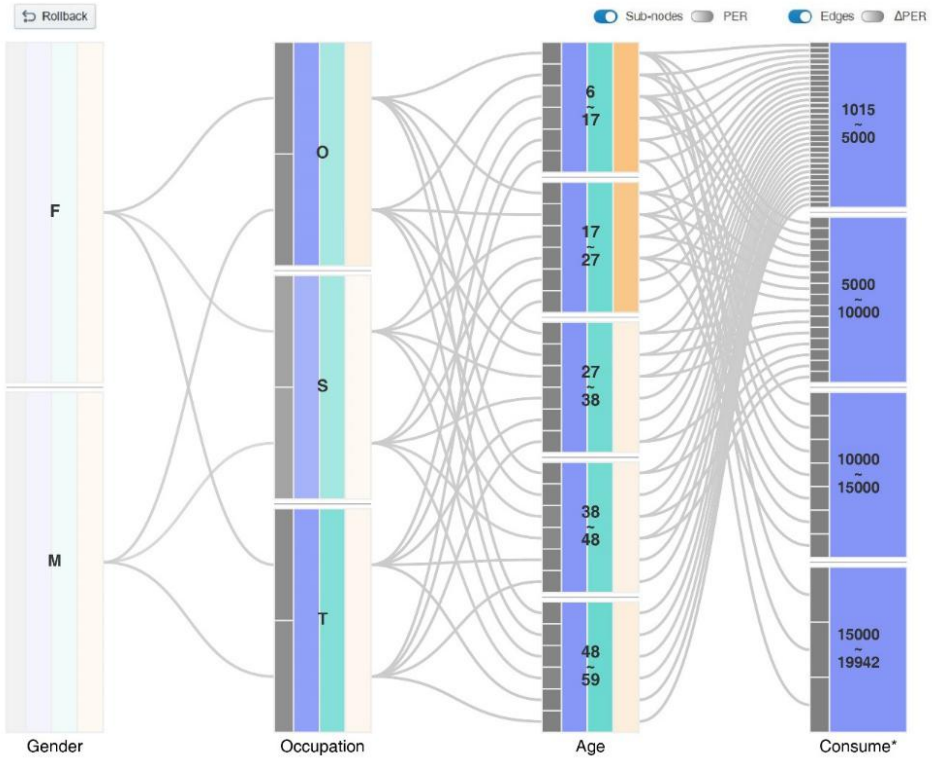
可视分析策略

VIS 2017

构建 PER-Tree

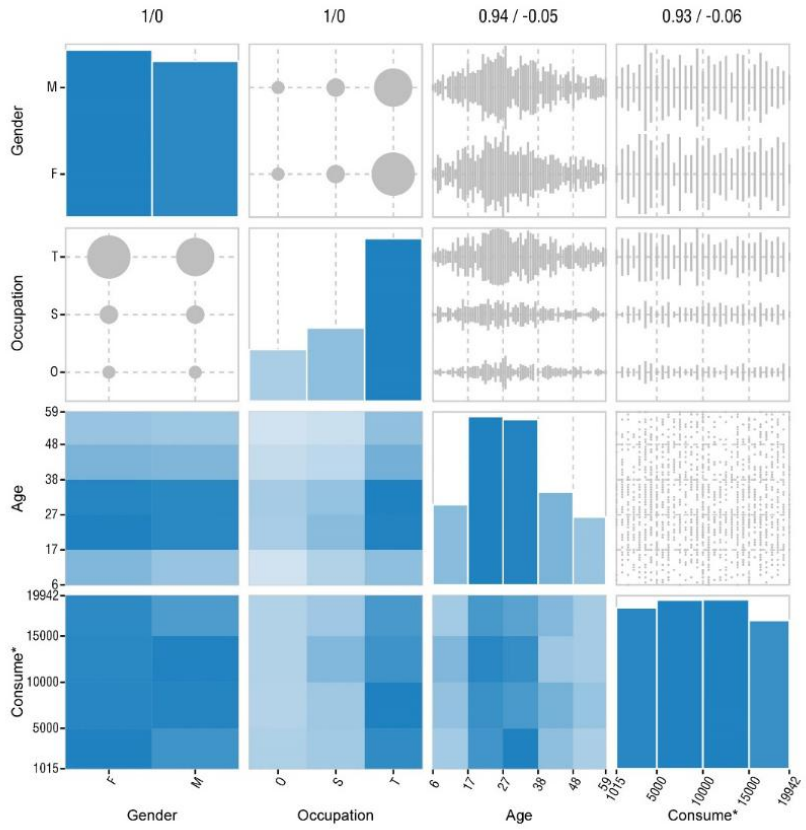


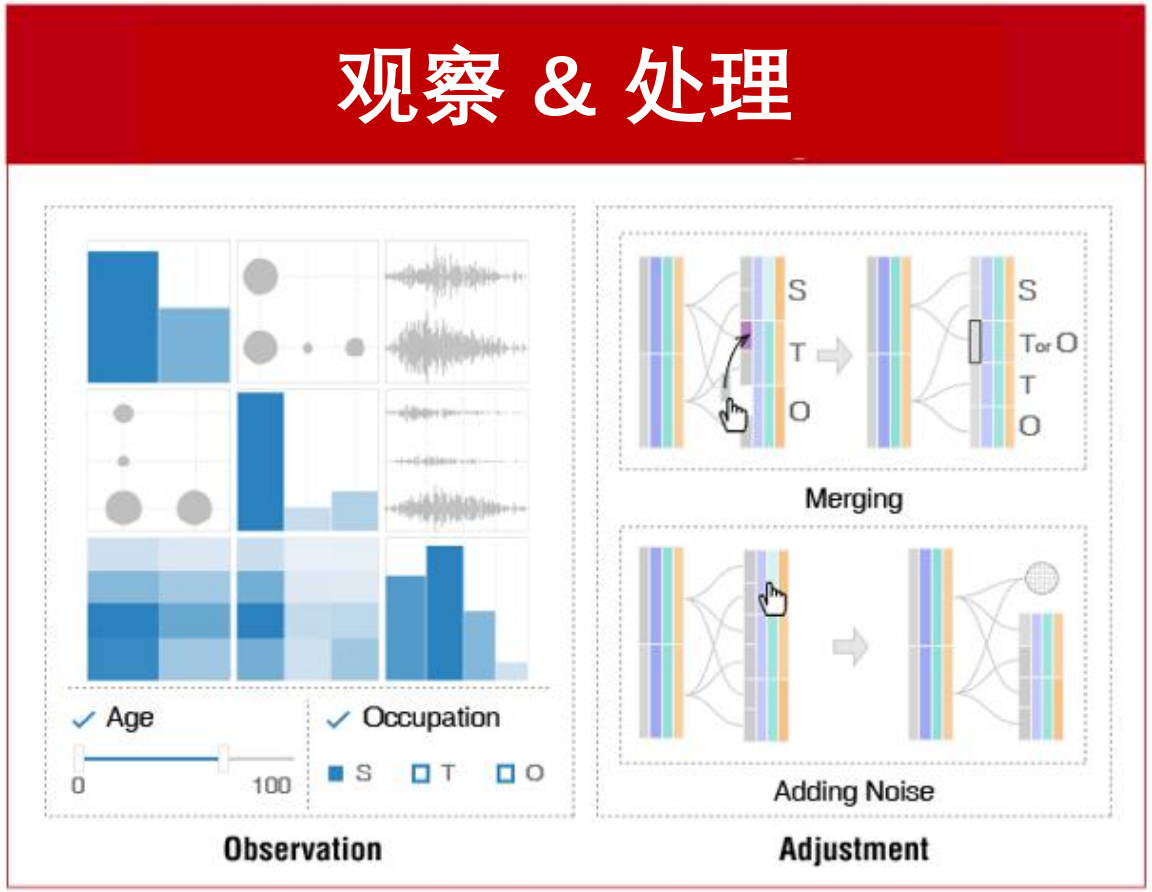
隐私暴露风险树
(Privacy Exposure Risk Tree)



观察 & 处理

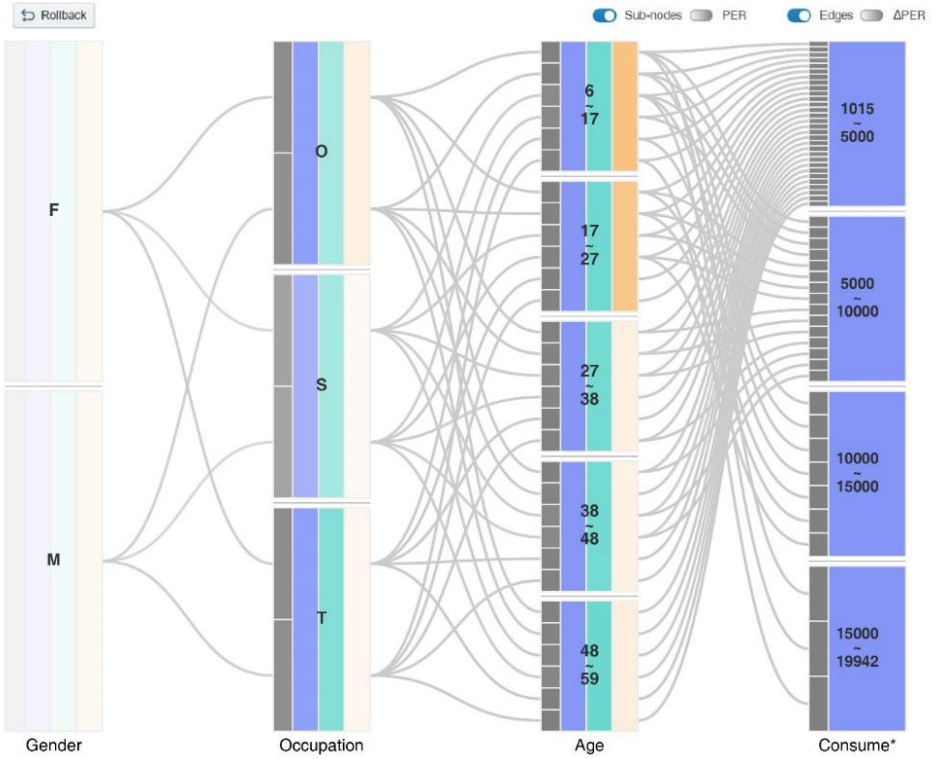
实用性度量矩阵
(Utility Preservation Degree Matrix)





PER-Tree

隐私暴露风险树

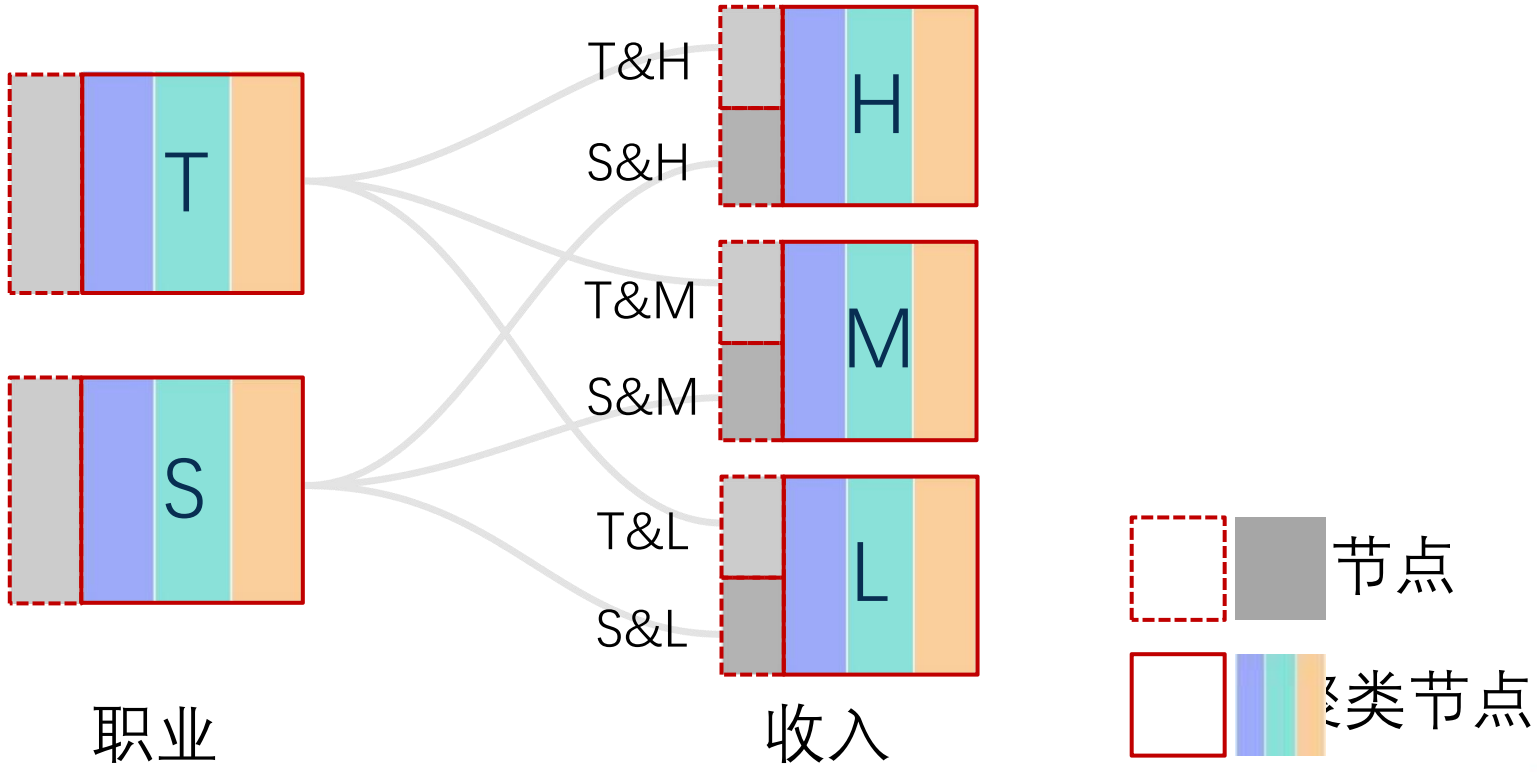


- **表达**表格数据
- **高亮**隐私问题
- **解决**隐私问题

VIS 2017

表达表格数据

职业	老师（T）
	学生（S）



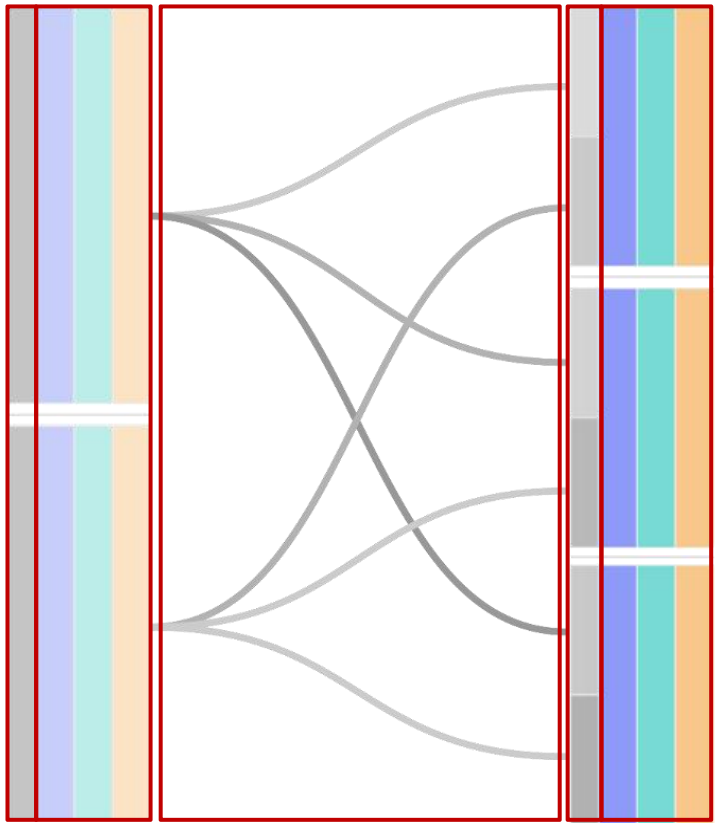
高亮隐私风险

子节点

隐私暴露风险

边

隐私暴露风险变动



聚类节点

个体数量 (k-anonymity)

属性值数量 (l-diversity)

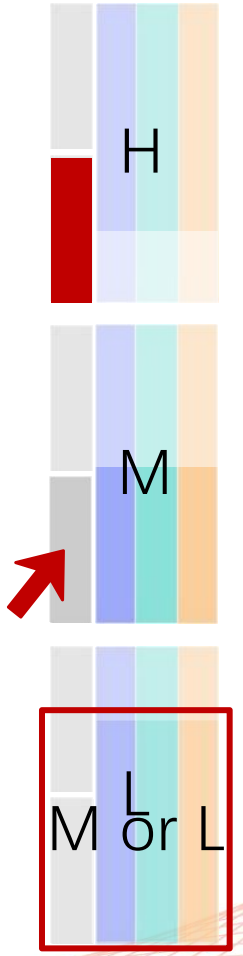
分布相似度 (t-closeness)

解决隐私问题

合并节点

模糊属性值

- 合并子节点 → 可能增加聚类节点
- 合并聚类节点 → 减少聚类节点



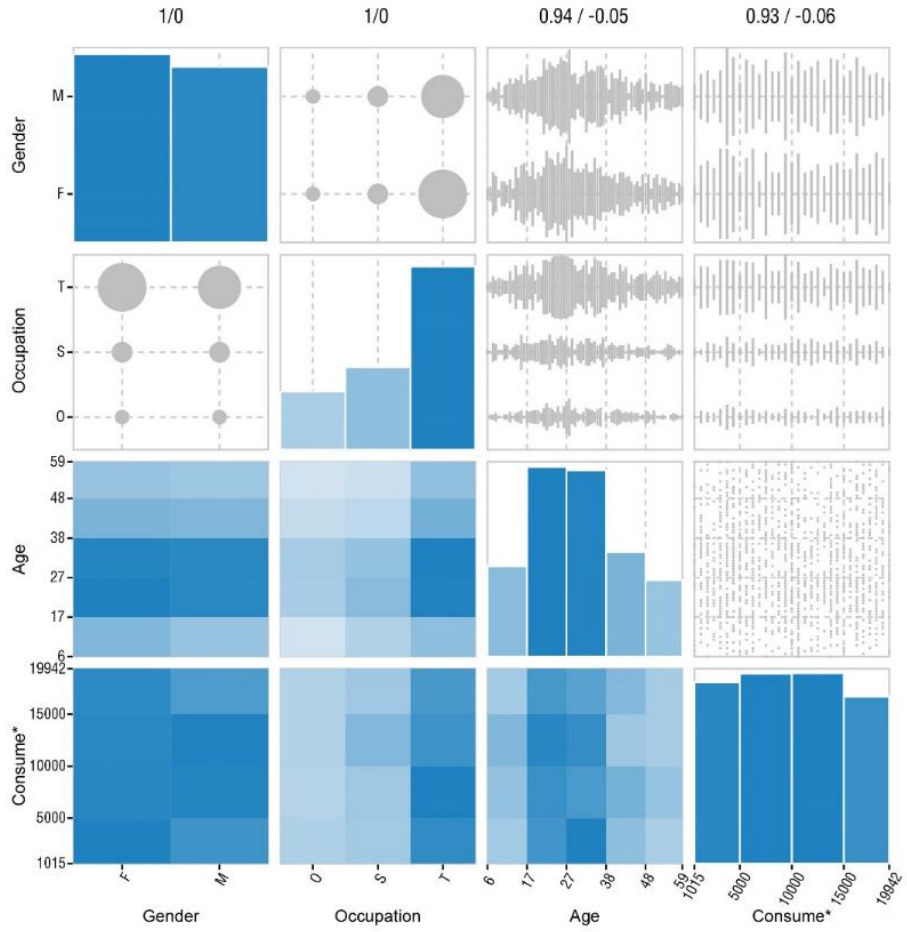
增加噪音

用噪音掩盖真实值

- 选择一组数据
- 选择一些属性
- 控制噪音的大小

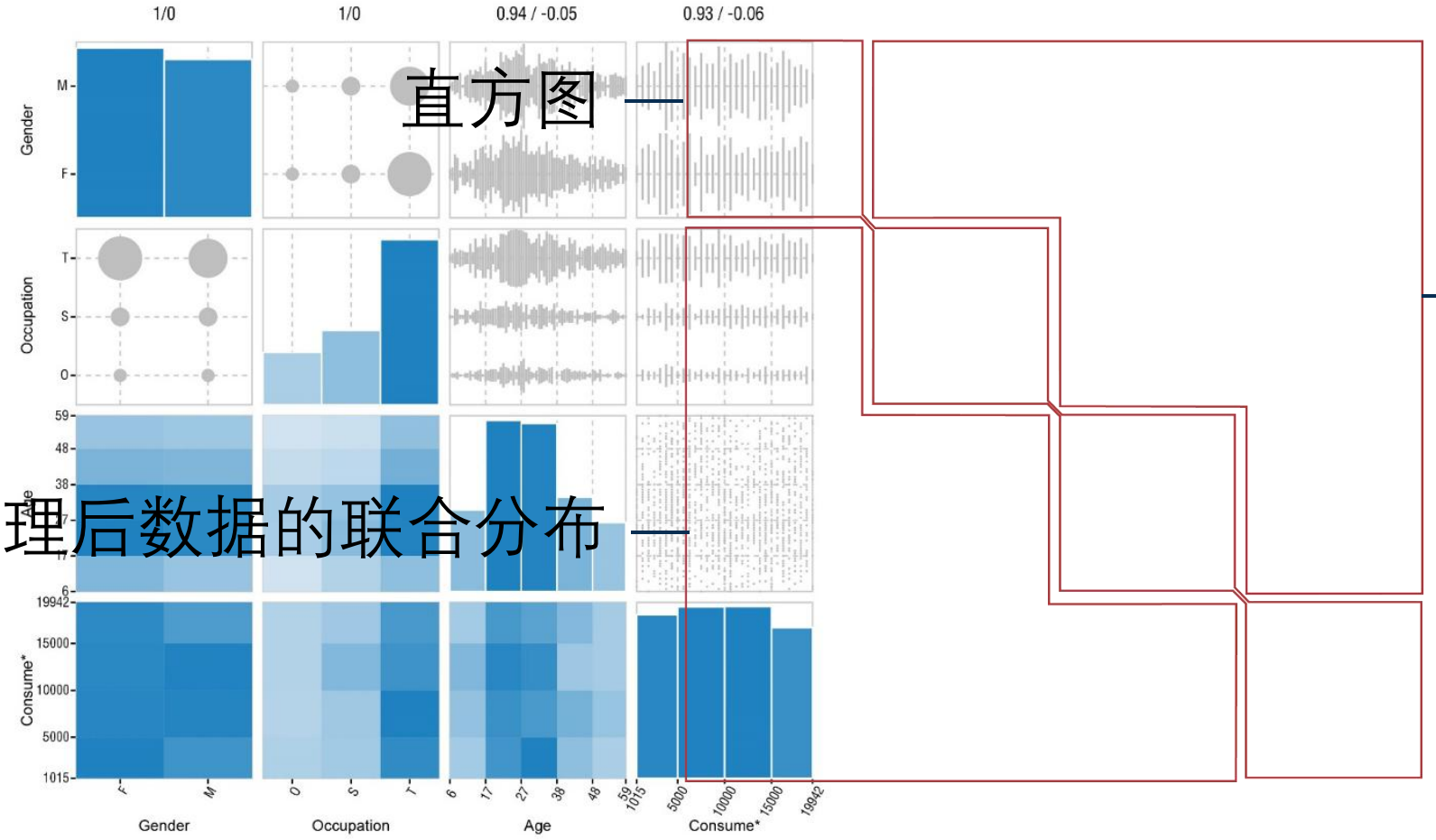
UPD-Matrix

实用性度量矩阵



- **表达**表格数据
- **评估**实用性

表达数据



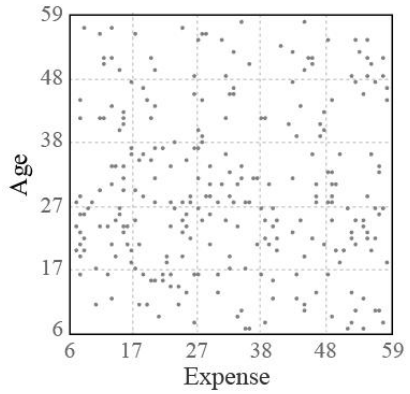
直方图

处理后数据的联合分布

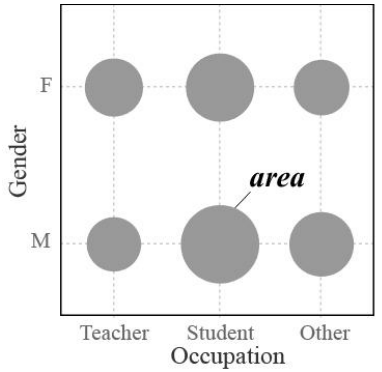
未处理数据的联合分布

联合分布

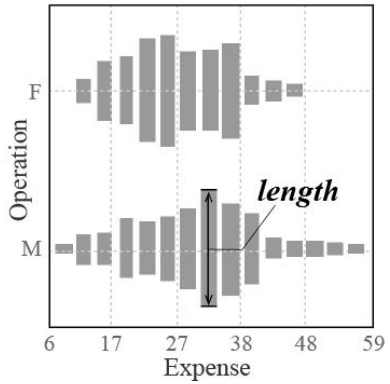
数值×数值



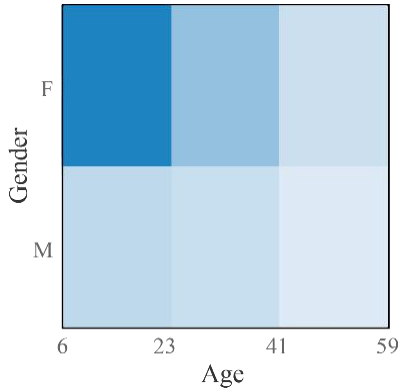
类别×类别



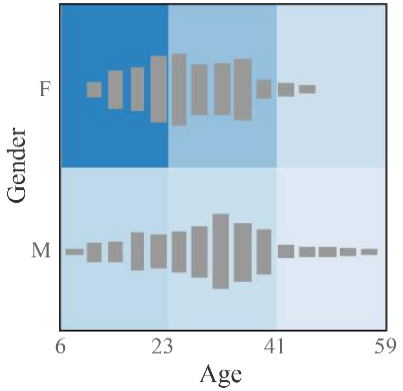
类别×数值



合并处理



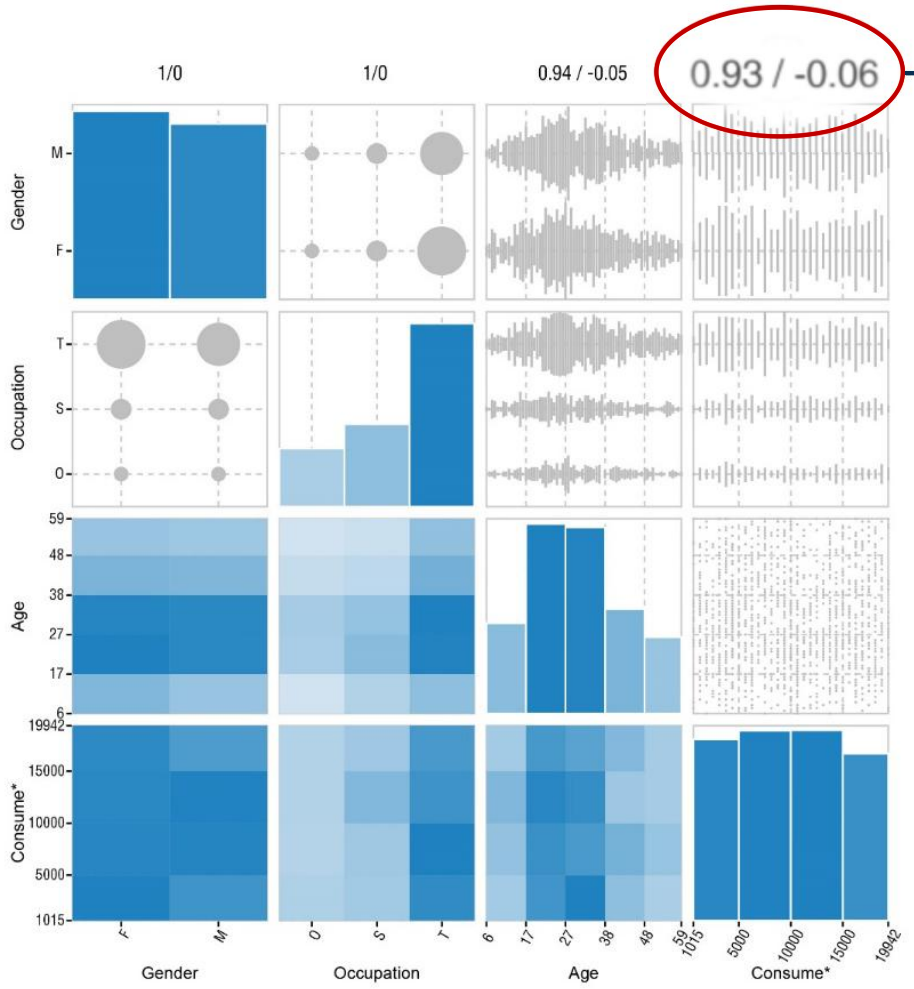
综合处理



未处理数据

处理后数据

评估实用性

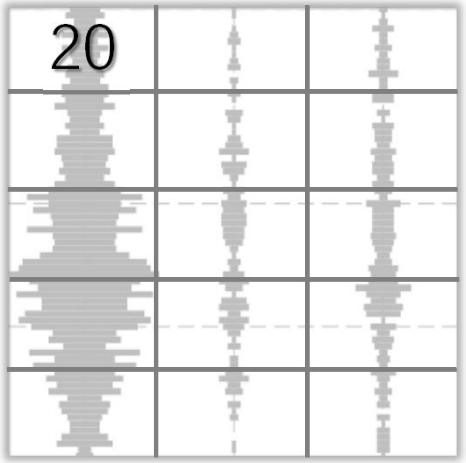


通过测地距离计算的实用性量化指标

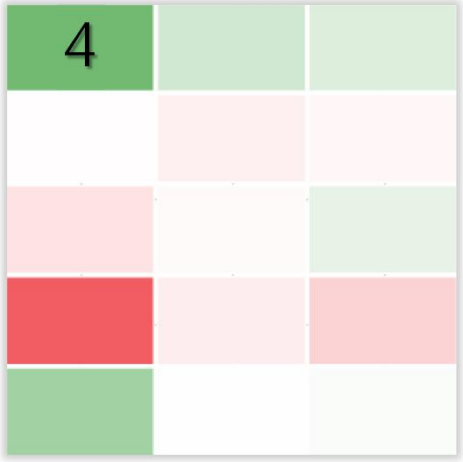
评估实用性

差别图

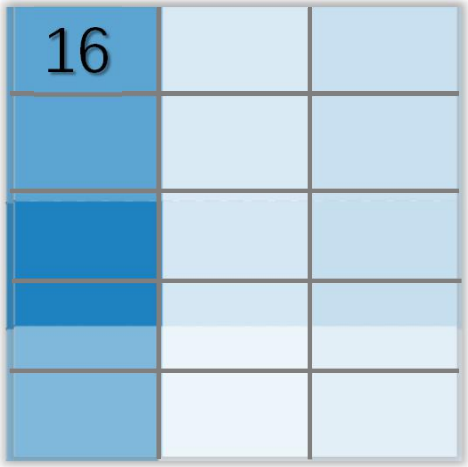
未处理数据



差异



处理后数据



正 + - 负

可视分析策略

VIS 2017

案例1：居民保险数据

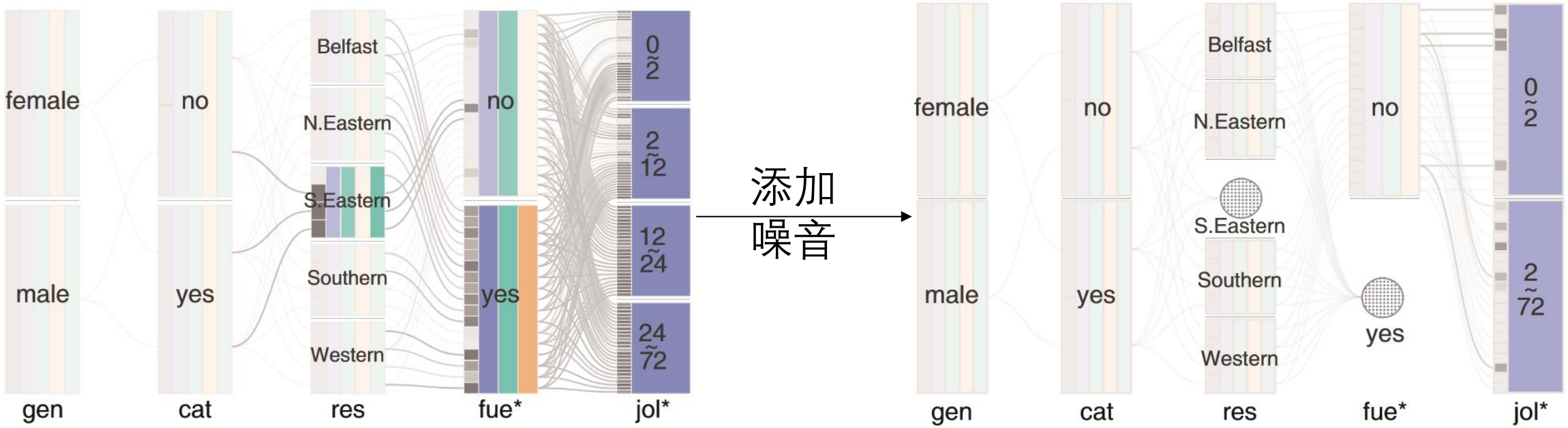
- 2015年美国怀俄明州的PUMS数据
- 1233条记录
- 4个必要属性：家庭成员中老人和孩子的个数、家庭收入以及保险支出
- 1个敏感属性：家庭收入

Case Study: Household Income and Insurance Census Data

案例2：毕业生发展数据

- 毕业生六年内的职业发展数据
- 712条记录
- 5个必要属性：性别、宗教信仰、居住地、父亲是否失业以及毕业多久入职
- 2个敏感属性：父亲是否失业以及 毕业多久入职

案例2：毕业生数据



只有一个聚类节点需要处理

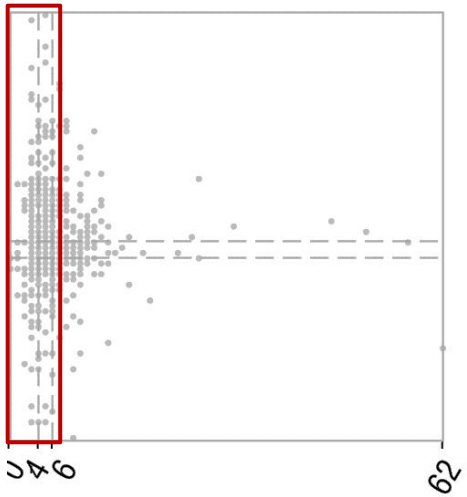
处理后的结果

案例3：服装业数据

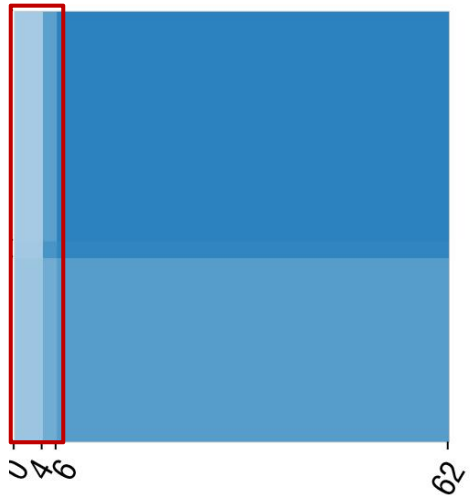
- 2004年中国企业数据
- 740条记录
- 4个必要属性：地区、商品类别、年利润、平均工资
- 1个敏感属性：平均工资

案例3：服装业数据

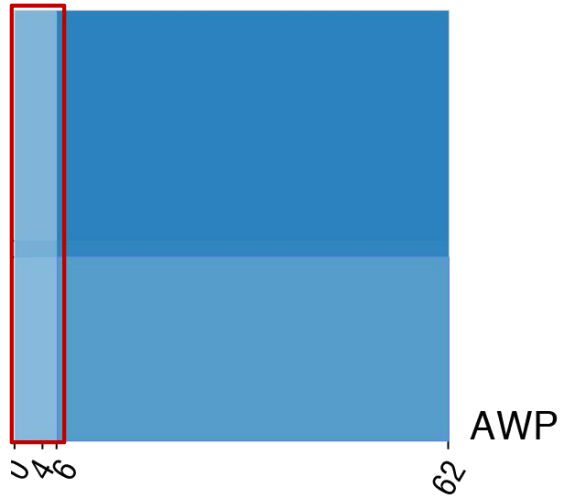
- 分组越多结果一定越好？ 不一定



原始数据分布



根据分布预分组



再次合并几乎没有
造成实用性损失

讨论

- 合并组或添加噪音
  - 语义匿名模型
  - 差分隐私模型

讨论

- 合并组或添加噪音
- 可视方法的灵活性
  - 哪种模型？
  - 应用于哪部分数据？
  - 施加多大的影响？

讨论

- 合并组或添加噪音
- 可视方法的灵活性
- 潜在局限
  - 实用性表达
  - 维度限制

谢谢


Xumeng Wang


Jia-Kai Chou


Wei Chen


Huihua Guan


Wenlong Chen


Tianyi Lao

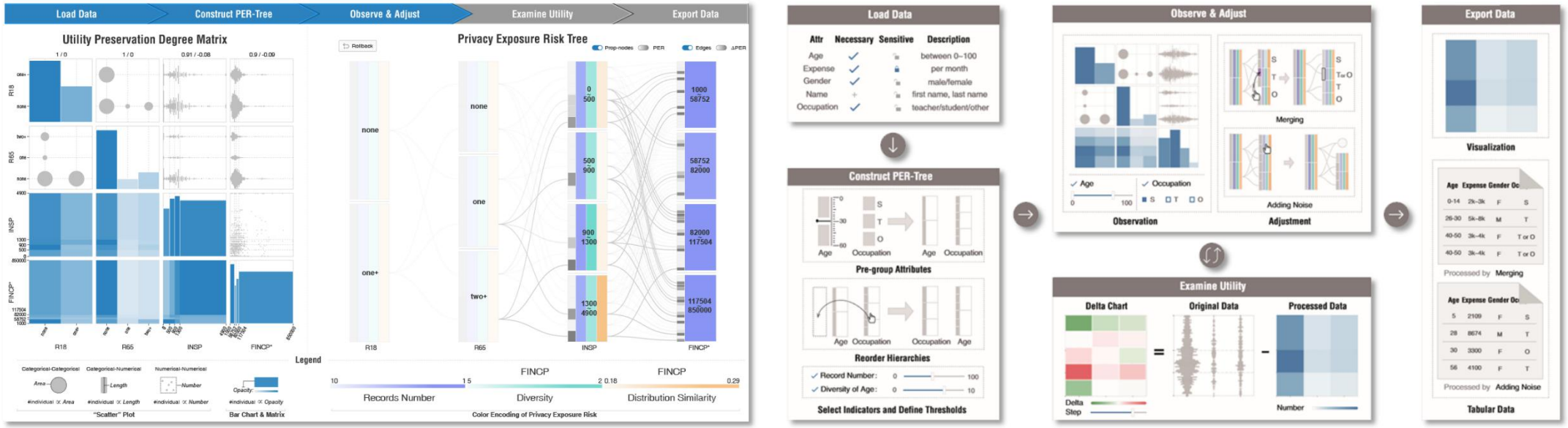

Kwan-Liu Ma

Acknowledgement

National 973 Program of China (2015CB352503)
National Natural Science Foundation of China (61232012 and 61422211)
U.S. National Science Foundation (IIS-1320229 and IIS-1528203)

Q&A

A Utility-aware Visual Approach for Anonymizing Multi-attribute Tabular Data



联系方式：wangxumeng@zju.edu.cn

VIS 2017