# Pattern Mining and Society: Privacy Issues

# Pattern Mining and Society: Privacy Issues

- A potential adverse side-effect of data mining—Privacy could be compromised
  - Privacy and accuracy are typically contradictory in nature
  - Improving one often incurs a cost on the other
- Three categories on privacy issues arising out of data mining
  - Input privacy (or data hiding)
    - Distort or hide data to prevent the miners from reliably extracting confidential or private information
  - Output privacy (or knowledge hiding)
    - No disclosure of sensitive patterns or knowledge from datasets
  - Owner privacy
    - Does not allow any party to reliably learn the data or sensitive information that the other owners hold (i.e., the source of the data)

# Ensuring Input Privacy

- Approach 1: Service provider anonymizes user's private information
  - B2B (business-to-business) environment
    - Service provider-to-data miner
  - Do you really trust them?
- Approach 2: Data anonymized/perturbed at the data source itself
  - B2C (business-to-customer) environment: Anonymized likely by a 3rd-party vendor
    - Methods: Data perturbation, transformation, or hiding (hide sensitive attributes)
- K-anonymity privacy requirement and subsequent studies
  - *k-anonymity*: Each equivalent class contains at least *k* records
  - It is still not sufficient, thus leads to further studies, such as *ℓ-diversity*, *t-closeness*, and *differential privacy*

| ID | ZIP | Age | Disease |
|---|---|---|---|
| 1 | 61801 | 45 | Heart |
| 2 | 61848 | 49 | Cancer |
| 3 | 61815 | 41 | Flu |
| 4 | 61804 | 32 | Diabetes |
| 5 | 61802 | 38 | Diabetes |
| 6 | 61808 | 39 | Flu |

| ID | ZIP | Age | Disease |
|---|---|---|---|
| 1 | 618** | 4* | Heart |
| 2 | 618** | 4* | Cancer |
| 3 | 618** | 4* | Flu |
| 4 | 618** | 3* | Diabetes |
| 5 | 618** | 3* | Diabetes |
| 6 | 618** | 3* | Flu |

# Data Perturbation for Privacy-Preserving Pattern Mining

- Statistical distortion: Using randomization algorithms
  - Independent attribute perturbation: Values in each attribute perturbed independently
  - Dependent attribute perturbation: Take care of correlations across attributes
- MASK [Rizvi & Haritsa VLDB'02]
  - Flip each 0/1 bit with a probability *p* (Note: this may increase a lot of items)
  - Tune *p* carefully to achieve acceptable average privacy and good accuracy
- Cut and paste (C&P) operator [Evfimievski et al. KDD'02]
  - Uniform randomization: Each existing item in the real transaction is, with a probability *p*, *replaced* with a new item not present in the original transaction
  - Methods developed on how to select items to improve the worst-case privacy
  - Experiments show mining a C&P randomized DB correctly identifies 80-90% of "short" (length ≤ 3) frequent patterns, but how to effectively mine long patterns remains an open problem

# Recommended Readings


- R. Agrawal and R. Srikant, Privacy-preserving data mining, SIGMOD'00
- C. C. Aggarwal and P. S. Yu, Privacy-Preserving Data Mining: Models and Algorithms, Springer, 2008
- C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science. 2014
- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In KDD'02
- A. Gkoulalas-Divanis, J. Haritsa and M. Kantarcioglu, Privacy in Association Rule Mining, in C. Aggarwal and J. Han (eds.), Frequent Pattern Mining, Springer, 2014 (Chapter 15)
- N. Li, T. Li, S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. ICDE'07
- A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, TKDD 2007
- S. Rizvi and J. Haritsa. Maintaining data privacy in association rule mining. VLDB'02
- J. Vaidya, C. W. Clifton and Y. M. Zhu, Privacy Preserving Data Mining, Springer, 2010