May 19-20, 2021
Faculty of Computer Engineering,
University of Isfahan

The 5th International Conference on Internet of Things and Its Application

A New Fast Framework for Anonymizing IoT Stream Data

University of Isfahan | Ferdowsi University of Mashhad

The 5th International Conference on Internet of Things and Its Application

AUTHORS

ALIREZA SADEGHI NASAB

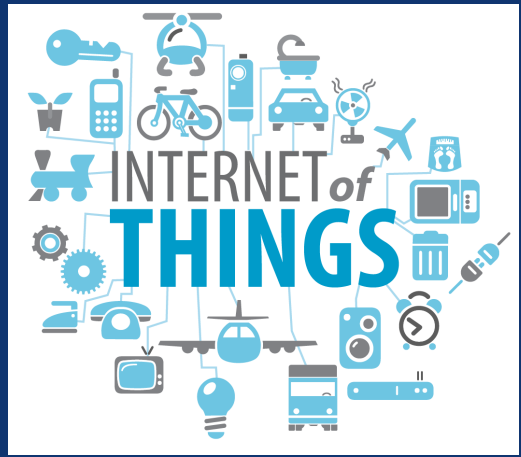
PhD candidate,
Arak University

HOSSEIN GHAFFARIAN

Assistant Professor,
Arak University

INTRODUCTION - IoT

- IoT becomes an integral part of modern life
- Some IoT applications in industries:
  - Energy consumption events of smart homes
  - Vital data on the activity of patients' organs
- To discover valuable knowledge to analyze user behavior, this data must be explored and mined [1]

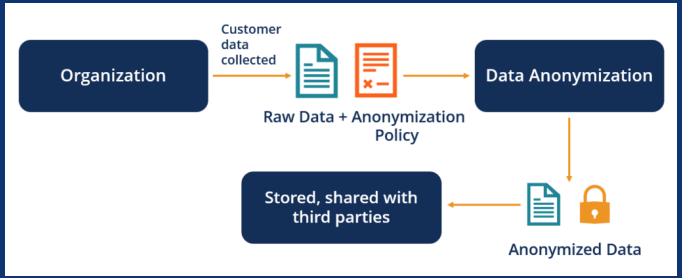


INTRODUCTION - DATA STREAMS

- Streaming data are widely used in today's world
- Some streaming data applications in industries:
  - Financial markets
  - Wireless sensors
  - Telecommunications
  - IoT, web applications, healthcare, etc [2].
- Extracting valuable knowledge from the streaming data can provide a realistic and approximate insight into indivisuals's activities. So, data anonymizing is important [3]

INTRODUCTION - PRIVACY PRESERVING

- One of the main concerns in the field of IoT is preserving users' privacy
- Anonymity is one of the most well-known methods used to protect user's personal data
- Anonymity is classifed into two main categories:
  - Static data anonymization
  - Streaming data anonymization [4]
- Anonymization quality in data streams is measured by information loss and average delay [5]



INTRODUCTION - ATTRIBUTE CLASSIFICATION

- Key attribute
- Quasi-identifier attribute
- Sensitive attribute
- Non-sensitive attribute [6]

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

INTRODUCTION - K-ANNONYMITY METHOD

- If a tuple in a table has some quasi-identifier value, at least another k-1 row must have the same QID value
- Probability of identifyng a tuple in k-anonymous table will be a maximum of $\frac{1}{k}$ [7]

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

RELATED WORKS

RELATED WORKS - CASTLE

- *CASTLE* is able to anonymize data streams according to *k-anonymity* or *l-diversity*
- It clusters the tuples into anonymity groups using a cluster distance metric called *enlargement*
- *Enlargement* metric measures the difference between the amount of generalization of the cluster before and after the tuple is assigned to the cluster
- The objective of clustering is to put similar tuples together into the same quasi-identifer group so that data distortion due to generalization is as small as possible
- *CASTLE* tends to cause accumulation of tuples into a single big cluster which is then over-generated [10]

RELATED WORKS - FADS

- *FADS* is another clustering-based anonymization approach for data streams
- It fixes the every cluster size to k. It uses buffer window like *FAANST* but it supports both categorical and numerical attributes
- Incoming tuples are accumulated within a buffer of a fixed length. Whenever the buffer gets full, an anonymization step is started which creates a cluster of tuples
- Oldest tuple plays the role of the seed member
- *FADS* is very efficient and yet effective method in terms of average information loss [13]

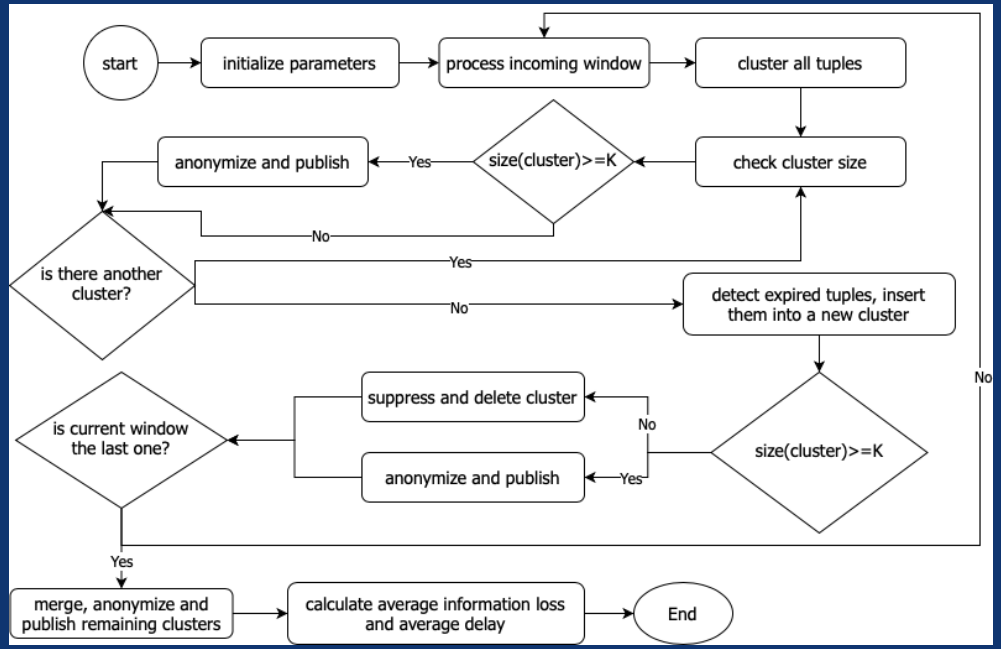
RELATED WORKS - UBDSA

- *UBDSA* is a tunable data stream k-anonymization framework
- The authors' motivation is to minimize average delay while keeping data quality high
- Data utility is a function of both data quality and data aging in data streams
- To attain high quality anonymity groups, It introduces a new distance metric, named *CAIL* (Cardinality Aware Information Loss)
- The algorithm dynamically balances the average information loss and data average delay by updating its parameters at runtime [15]

RELATED WORKS - Overall Comparison

Algorithm	Pros	Cons
CASTLE	Supporting l-diversity	Possibility to create superclusters
FADS	Low information loss	High average delay
UBDSA	Able to tune and balance information loss and average delay	High complexity of merge and split functions

PROPOSED METHOD - FLOWCHART



Implementation: Scala [16] and Apache Flink [17]

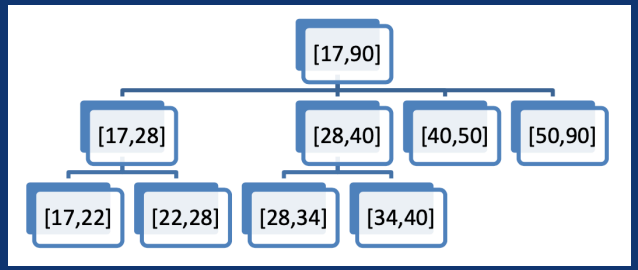
Scala

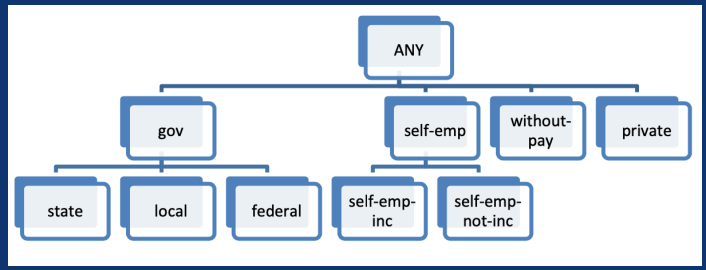
Apache Flink

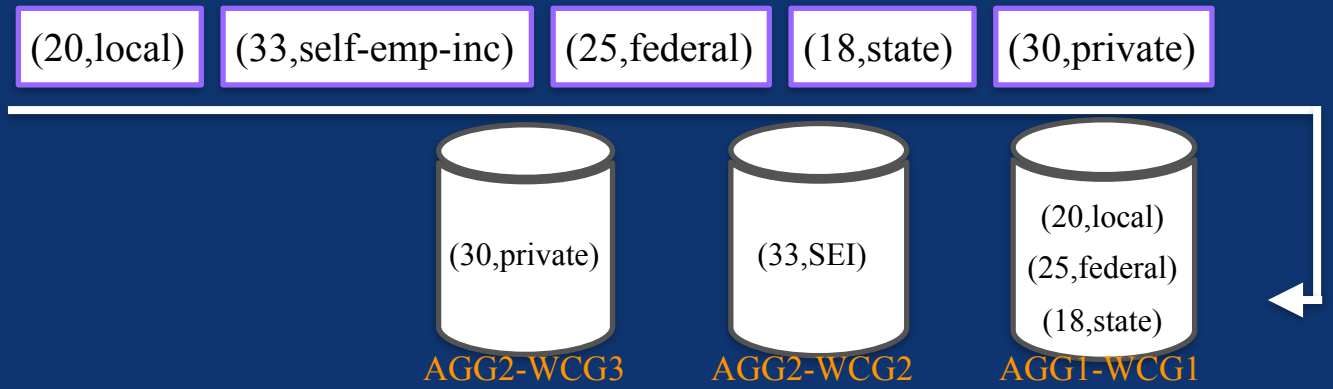
A New Fast Framework for Anonymizing IoT Stream Data

The 5th International Conference on Internet of Things and Its Application, Faculty of Computer Engineering, University of Isfahan.

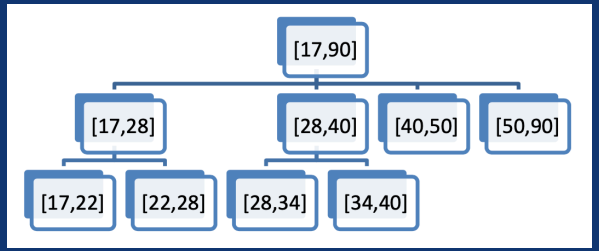
PROPOSED METHOD - NEW CLUSTERING METHOD

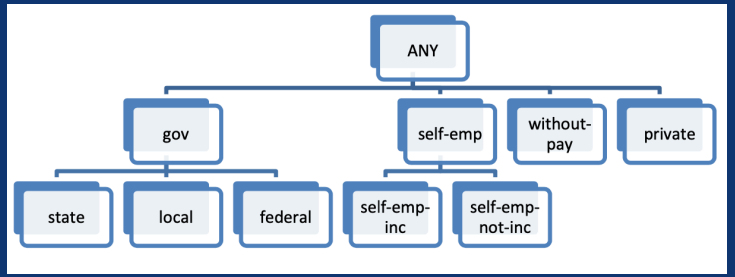






PROPOSED METHOD - GENERALIZATION





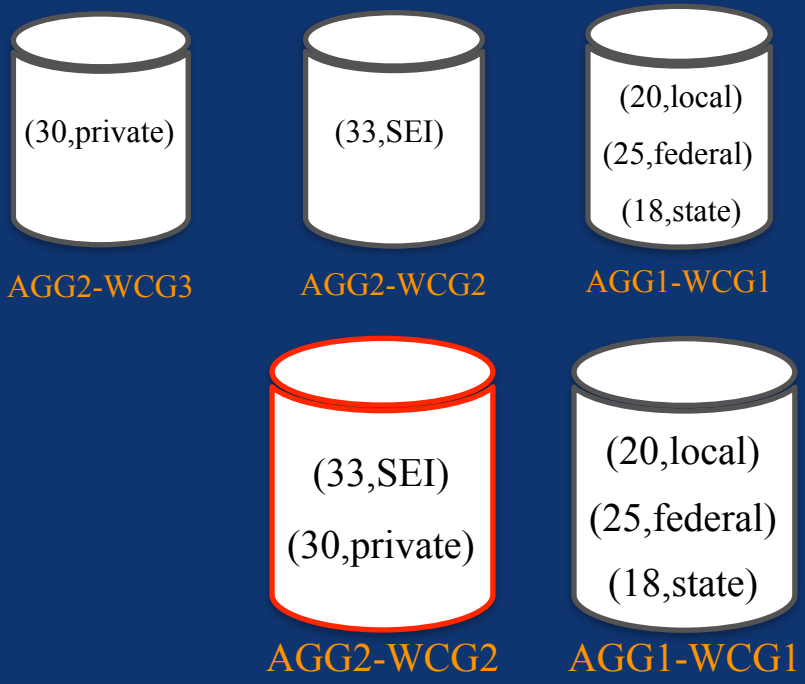
$$IL_{age} = \frac{2-1}{6-1} = 0.2$$

$$IL_{workClass} = \frac{3-1}{7-1} = \frac{1}{3}$$

A New Fast Framework for Anonymizing IoT Stream Data

The 5th International Conference on Internet of Things and Its Application,
Faculty of Computer Engineering, University of Isfahan.

PROPOSED METHOD - MERGE REMAINING CLUSTERS



RESULTS - EVALUATION SPECIFICATIONS

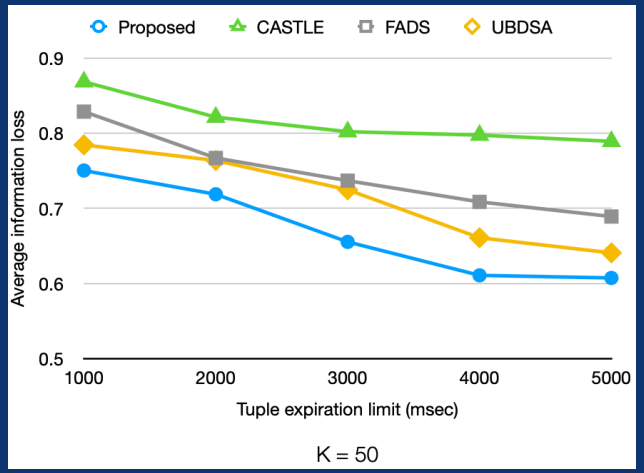
- We calculate average information loss and average delay for evaluation
- Results are compared with three researches: *CASTLE*, *FADS* and *UBDSA*
- Expriments are conducted on a computer with 4 VCPs and 15 GB of RAM, running in Ubuntu linux operation system
- Adult dataset [18] has been used to evaluate proposed method which is very famous dataset in data anonymization field

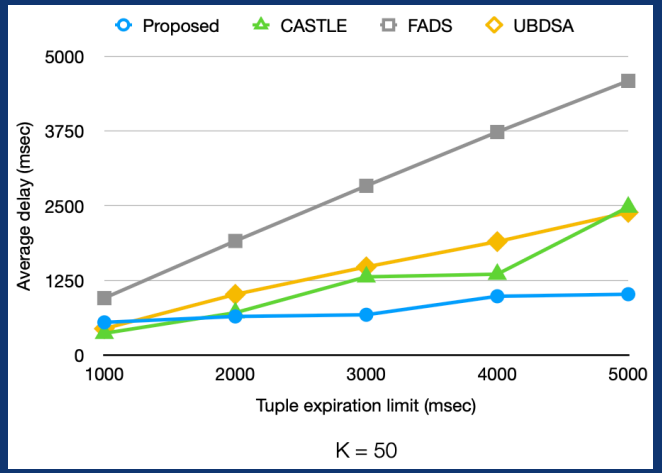
RESULTS - ADULT DATASET SPECIFICATIONS

Attribute Name	Attribute Data Type	Attribute Type
Age	Quasi-identifier	Numerical
Work class	Quasi-identifier	Categorical
Final weight	Quasi-identifier	Numerical
Education number	Quasi-identifier	Numerical
Education	Quasi-identifier	Categorical
Marital status	Quasi-identifier	Categorical
Nation	Sensitive	Categorical

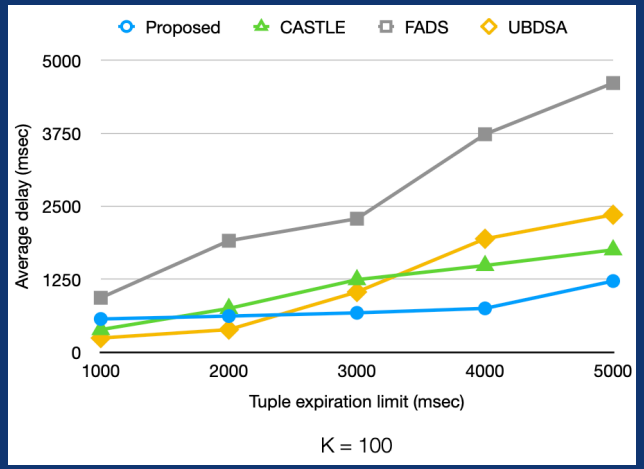
Attribute Name	Attribute Data Type	Attribute Type
Occupation	Quasi-identifier	Categorical
Relationship	Quasi-identifier	Categorical
Race	Quasi-identifier	Categorical
Gender	Quasi-identifier	Categorical
Capital gain	Quasi-identifier	Numerical
Capital loss	Quasi-identifier	Numerical
Hours per week	Quasi-identifier	Numerical
Income	Quasi-identifier	Categorical

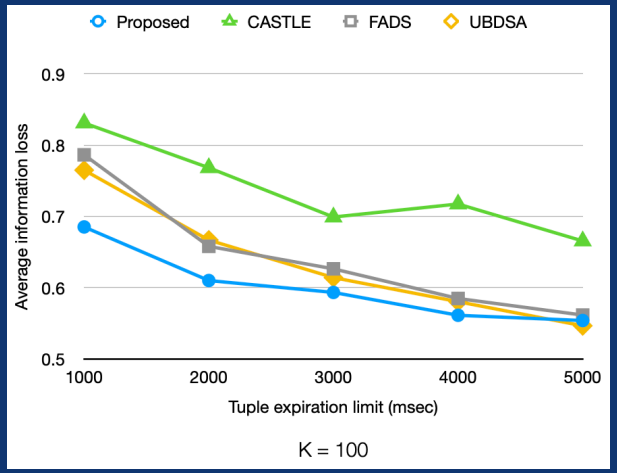
RESULTS (K = 50)


Proposed
CASTLE
FADS
UBDSA
0.9
0.8
0.7
0.6
0.5
Average information loss
1000
2000
3000
4000
5000
Tuple expiration limit (msec)
K = 50


Proposed
CASTLE
FADS
UBDSA
5000
3750
2500
1250
0
Average delay (msec)
1000
2000
3000
4000
5000
Tuple expiration limit (msec)
K = 50

RESULTS (K = 100)





CONCLUSION

- Data explosion is a double-edged sword. With the advancement of technology, privacy has become one of concerning areas nowadays
- The proposed method works better than the others by considering two key components in the anonymity of stream data, i.e average information loss and average data delay
- Innovations of proposed method:
  - Introducing new clustering method based on class labels
  - Introducing new method for classifying and anonymizing numerical data based on attribute statistics information
  - Use of data processing engine which is suitable for streaming data

REFERENCES

1. A. Otgonbayar, Z. Pervez and K. Dahal, “Toward anonymizing IoT data streams via partitioning,” *IEEE* 13*th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* ,Brasilia, 2016.
2. Y. Tian, J. Yuan and Y. Hou, “PDF-DS: Privacy-Preserving Data Filtering for Distributed Data Streams in Cloud,” International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) ,Atlanta, GA, USA, 2019.
3. S. Kim, M. K. Sung and Y. D. Chung, “A framework to preserve the privacy of electronic health data streams,” *Journal of biomedical informatics*, vol. 50, pp. 95-106, 2014.
4. B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao and Y. Jia, “Continuous privacy-preserving publishing of data streams,” 12*th International Conference on Extending Database Technology: Advances in Database Technology* ,Saint Petersburg, Russia, .2009
5. A. Otgonbayar, Z. Pervez, K. Dahal and S. Eager, “K-VARP: K- anonymity for varied data streams via partitioning,” *Information Sciences*, vol. 467, p. 238-255, 2018.
6. B. C. Fung, K. Wang, A. W.-C. Fu and S. Y. Philip, Introduction to privacy-preserving data publishing: Concepts and Techniques, New York, USA: CRC Press, 2010.
7. L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

REFERENCES

8. W.Wang, J.Li, C.Ai and Y.Li, “Privacy protection on sliding window of data streams,” *International Conference on Collaborative* Computing: Networking, Applications and Worksharing (CollaborateCom 2007) ,New York, USA, 2007.
9. J. Li, B. C. Ooi and W. Wang, “Anonymizing streaming data for privacy protection,” *IEEE 24th International Conference on Data Engineering* , Cancun, Mexico, 2008.
10. J. Cao, B. Carminati, E. Ferrari and K.-L. Tan, “Castle: Continuously anonymizing data streams,” EEE Transactions on Dependable and Secure Computing, vol. 8, no. 3, pp. 337-352, 2010.
11. H. Zakerzadeh and S. L. Osborn, “Faanst: fast anonymizing algorithm for numerical streaming data,” *Data privacy management and autonomous spontaneous security* ,Berlin, Heidelberg, Springer, pp. 36- 50, 2010.
12. P. Wang, J. Lu, L. Zhao and J. Yang, “B-castle: An efficient publishing algorithm for k-anonymizing data streams,” *Second WRI Global Congress on Intelligent Systems* ,Wuhan, 2010.
13. K. Guo and Q. Zhang, “Fast clustering-based anonymization approaches with time constraints for data streams,” *Knowledge-Based Systems*, vol. 46, pp. 95-108, 2013.
14. E. Mohammadian, M. Noferesti and R. Jalili, “FAST: fast anonymization of big data streams,” *International Conference on Big Data Science and Computing* ,Beijing, 2014.

REFERENCES

15. U. Sopaoglu and O. Abul, “A utility-based approach for data stream anonymization,” *Journal of Intelligent Information Systems*, pp. 1-27, 2019.
16. “The Scala Programming Language,” Programming Methods Laboratory of École polytechnique fédérale de Lausanne, .Available: https://www.scala-lang.org.
17. “Apache Flink - Stateful Computations over Data Streams,” Apache Software Foundation, Available: https://flink.apache.org.
18. “Adult,” Uci machine learning repository, Available: https://archive.ics.uci.edu/ml/datasets/adult.

Thanks

Address: Computer Engineering Faculty, University of Isfahan, Hezar Jarib Ave, Isfahan, Iran.

Telephone - Fax: +98 - 31 - 3793 5642

E-mail: iot2021@ res.ui.ac.ir