Effective and Efficient Anonymization of Health-Related Physical Activity Data

Pooja Parameshwarappa, Zhiyuan Chen and Gunes Koru

University of Maryland, Baltimore County

Motivation - I

- Availability of physical activity data is increasing
  - Use of wearable devices [Kaewkannate et al., 2016]
  - Use of smart-phone fitness applications [Higgins, J.P., 2016]
  - Data from smart-environments [Skocir et al., 2016 and Jakkula et al., 2011]

- Publishing activity data can help
  - Research in fighting chronic diseases [Ermes et al., 2008]
  - Research in reducing health care costs [Spenkelink et al., 2002, Centers for Disease Control and Prevention, 2015]
  - Reproducible research [Peng, 2011]
  - Facilitate teaching data analytics

- However, publishing activity data comes with high privacy risks
  - Re-identification

Motivation – II: Re-identification Example

- Adversary's knowledge
  - Victim's record is in the data set
  - The victim runs at 6:00 am every Monday, Wednesday, and Friday

6 am Monday | 6 am Wednesday | 6 am Monday

Activity Data	Disease
SSS...WSS...SSS......SSS...SSS...SSS...	Heart-disease
SSS...SSW...SSS......SSS...SSS...SSS...	Cold
RSS...RSW...RSS......RSS...RSS...RSS...	Depression
SSS...SSS...SSS......SSS...SSS...SSS...	Heart-disease

- Possible Solution: Anonymization

Background (k-Anonymity)

- A given data set is said to satisfy k-anonymity if every record in a table has at least k-1 other records that are identical with respect to the quasi-identifiers [Sweeney, 2002]

Age	Sex	Zipcode	Disease
22	M	21220	Cold
25	M	21222	Heart-disease
33	M	21235	Cancer
30	F	21232	Cancer
28	F	21234	Cancer

Original Data

Age	Sex	Zipcode	Disease
[22 – 25]	*	2122*	Cold
[22 – 25]	*	2122*	Heart-disease
[28 – 33]	*	2123*	Cancer
[28 – 33]	*	2123*	Cancer
[28 – 33]	*	2123*	Cancer

2-Anonymous Data

Limitation of Existing Anonymization Techniques: Dealing With Sequential Data

- Most of the anonymization techniques are suitable for cross-sectional data sets
- Activity data is sequential in nature
  - Each time point acts a dimension
  - Very high dimensionality

1 2 3

Age	Sex	Zipcode	Disease
22	M	21220	Cold
25	M	21222	Heart-disease
33	M	21235	Cancer
30	F	21232	Cancer
28	F	21234	Cancer

Cross-sectional Data

123... n

Activity Data
SSS...WSS...SSS......SSS...SSS...SSS...
SSS...SSW...SSS......SSS...SSS...SSS...
RSS...RSW...RSS......RSS...RSS...RSS...
SSS...SSS...SSS......SSS...SSS...SSS...

Sequential Data

Limitations of Existing Anonymization Techniques

- For sequential data, entire sequence represents the potential quasi-identifiers

Age	Sex	Zipcode	Disease
22	M	21220	Cold
25	M	21222	Heart-disease
33	M	21235	Cancer
30	F	21232	Cancer
28	F	21234	Cancer

Cross-sectional Data

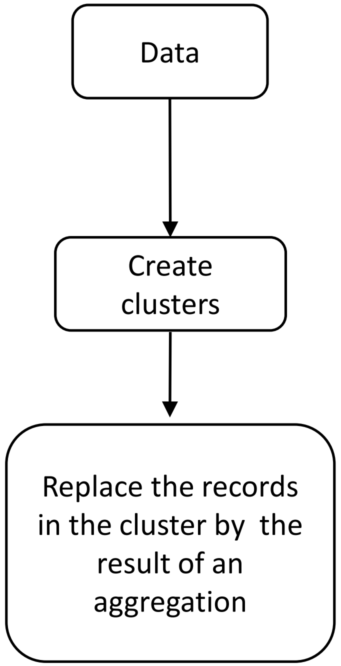
?

Activity Data
SSS...WSS...SSS......SSS...SSS...SSS...
SSS...SSW...SSS......SSS...SSS...SSS...
RSS...RSW...RSS......RSS...RSS...RSS...
SSS...SSS...SSS......SSS...SSS...SSS...

Sequential Data

Method

- One of the approaches to achieve k-anonymity is Microaggregation [Domingo-Ferrer et al., 2006]



Age	Sex	Zipcode	Disease
22	M	21220	Cold
25	M	21222	Heart-disease
33	M	21235	Cancer
30	F	21232	Cancer
28	F	21234	Cancer

22	M	21220	Cold
25	M	21222	Heart-disease

33	M	21235	Cancer
30	F	21232	Cancer
28	F	21234	Cancer

22	M	21220	Cold
22	M	21220	Heart-disease

30	F	21232	Cancer
30	F	21232	Cancer
30	F	21232	Cancer

- A well-known heuristic method for achieving microaggregation is, Maximum Distance to Average Vector (MDAV) [Domingo-Ferrer et al., 2006, Solanas et al., 2006]

Proposed Approach

Step 1: Multi-level Clustering (MC)

- At the root level, all the activity sequences are assigned to one cluster
- In the subsequent levels:
  - Sequences are aggregated to certain time intervals (dimensionality reduction)
  - Clustered using MDAV
- This process is repeated until each cluster at the leaf-level has at least k sequences

Step 2: Anonymization (MCKA – Multi-level Clustering Based K-Anonymity)

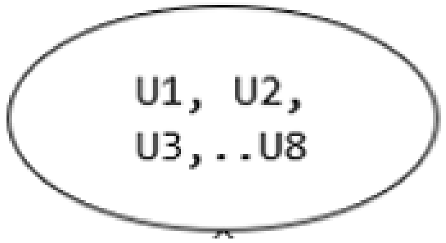
- K-Anonymity is applied to each leaf-level cluster

Proposed Approach - Example

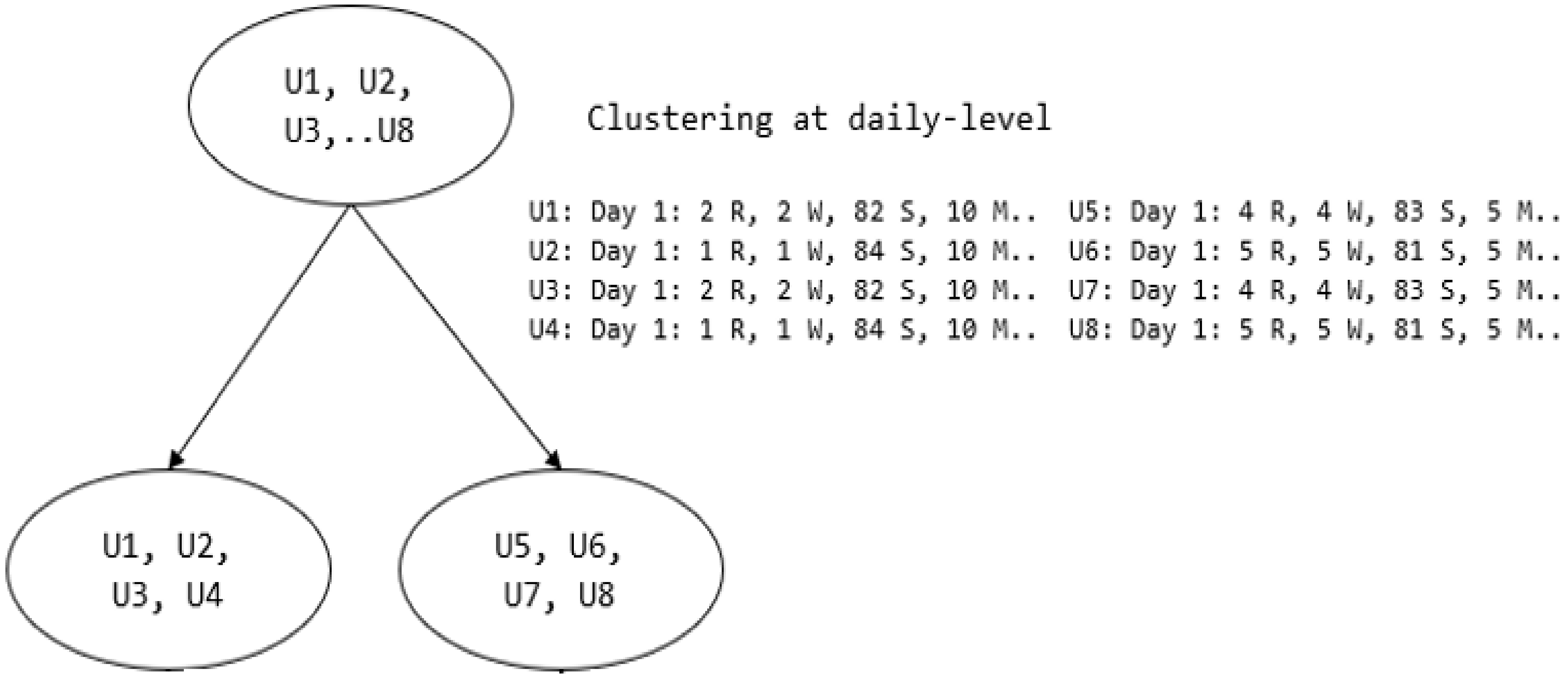
Step 1: Multi-level Clustering (MC)

- At the root level, all the sequences are assigned to one cluster

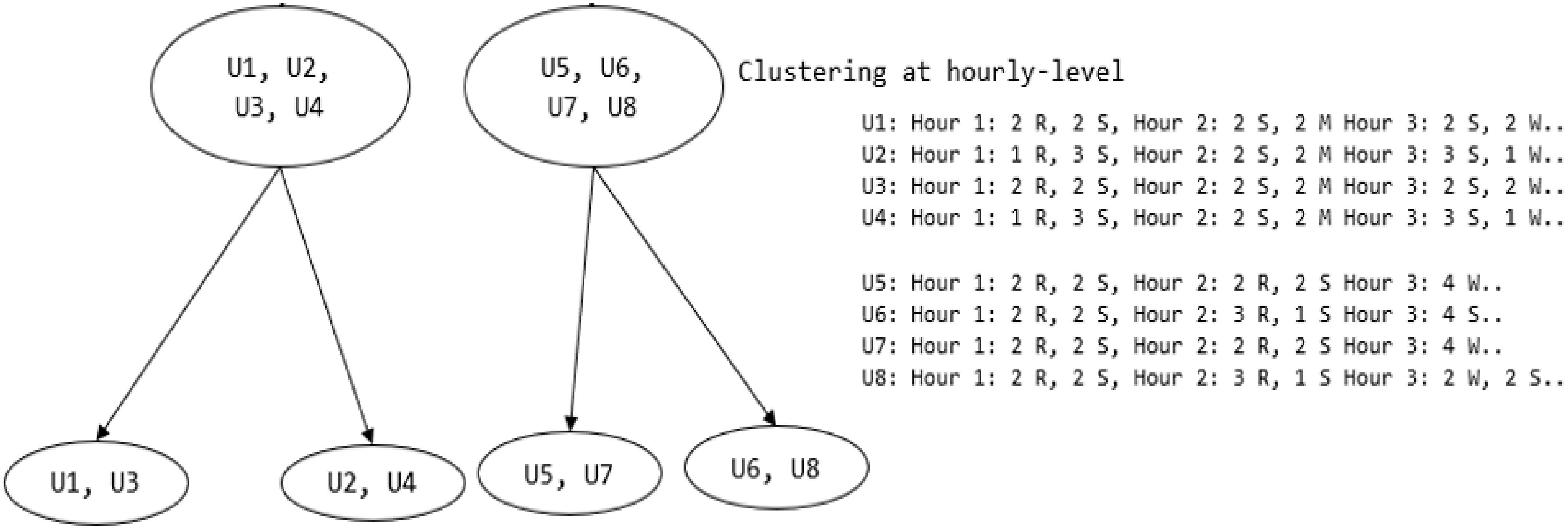
```
U1: S, S, R, R, S, S, M, M, S, S, W, W, ..
U2: S, S, R, S, S, S, M, M, S, S, W, S, ..
U3: R, R, S, S, M, M, S, S, W, W, S, S, ..
U4: R, S, S, S, M, M, S, S, S, W, S, S, ..
U5: S, S, R, R, S, S, R, R, W, W, W, W, ..
U6: S, S, R, R, S, R, R, R, S, S, S, S, ..
U7: S, S, R, R, R, R, S, S, W, W, W, W, ..
U8: S, S, R, R, S, R, R, R, S, S, W, W, ..
```



- The sequences are then aggregated to certain time intervals, and then clustered using MDAV



- In the next level, the sequences are drilled down to smaller time intervals and clustered using MDAV
- These steps are repeated until each cluster at the leaf level has at least k sequences



Step 2: Multi-level Clustering Based K-Anonymity (MCKA)

- Compute centroid for each cluster
- Instead of replacing the sequences in a cluster with the centroid, simulate as many sequences as the size of the cluster by using the centroid

U1	S	S	R	R	S	S	M	M	S	S	W	W	..
U2	S	S	R	S	S	S	M	M	S	S	W	S	..
U3	R	R	S	S	M	M	S	S	W	W	S	S	..
U4	R	S	S	S	M	M	S	S	S	W	S	S	..

Compute cluster center

S	0.5	0.75	0	0.75	0.5	0.5	0.5	0.5	0.75	0.5	0.5	0.75	..
R	0.5	0.25	0.5	0.25	0	0	0	0	0	0	0	0	..
W	0	0	0.5	0	0	0	0	0	0.25	0	0	0.25	..
M	0	0	0	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0	..

Generate anonymized data using
probabilistic sampling

U1	R	R	W	S	S	S	M	M	S	S	S	S	..
U2	S	S	W	R	M	M	S	S	S	M	M	S	..
U3	S	S	W	S	M	M	M	S	S	M	M	W	..
U4	S	R	R	S	S	S	S	S	W	S	S	S	..

Experimental Design - I

- Student Life data set was used [Wang et al., 2014]
- Original data had activity information (Stationary, Walking, Running and Missing) for 49 students
- Synthetic data was generated for 9800 students (1.9 GB)
- System configuration
  - 32 GB RAM, 3.3 GHz processor
  - Windows 10 operating system
  - R programming (Single thread implementation)

Experimental Design - II

- Activity information was available at minute-level intervals, for two weeks, for every student
- Activity sequences were clustered using MC (proposed approach) and using MDAV (conventional approach)
- Clusters were anonymized using k-anonymity
- Efficiency
  - Time taken for clustering
- Utility
  - Relative difference between un-anonymized and anonymized data
  - Correlations between activity and other attributes (Flourishing scale, CGPA)

Comparing MC and MDAV -- Efficiency

	MC (k=5)	MDAV (k=5)	MDAV (k=5)
		(Daily)	(No aggregation)
Time for clustering	21 mins	2.6 hrs	>12.6 hrs (memory issues)

- MDAV on the original data set ran out of memory
- Upon aggregation to higher time intervals, MDAV completed in 2.6 hours and MC completed in 21 mins

Comparing MCKA and MDAV-KA -- Utility

- Data was clustered using both MC and MDAV.
- K-Anonymity was applied on the resulting clusters
- Relative difference between un-anonymized data and anonymized data was computed

	MCKA k=5	MDAV-KA k=5
Daily (S)	0.08	0.08
Daily (W)	0.23	0.22
Daily (R)	0.17	0.16
Daily (M)	0.22	0.21

- MCKA and MDAV-KA showed comparable results

Preserving Correlations -- Utility

- Direction and magnitude of the correlations were preserved after anonymization

	Activity-Flourishing		Activity-CGPA	
	Correlation (r)	p-value	Correlation (r)	p-value
Un-anonymized data	0.146	$<$ 2.2e-16	-0.289	$<$ 2.2e-16
MCKA (k=5)	0.146	$<$ 2.2e-16	-0.290	$<$ 2.2e-16
MCDP (k=50)	0.129	$<$ 2.2e-16	-0.293	$<$ 2.2e-16

Discussion

- Efficiency
  - MC reduces computation time from hours to minutes
- Utility
  - MC-KA has relative difference comparable with MDAV-KA
  - Direction and magnitude of the correlations are preserved
- Privacy
  - MCKA guarantees k-Anonymity
  - Suitable for both small and large data sets

Conclusion

- The proposed approach preserves privacy and utility, and in addition, reduces the computation time from hours to minutes

- To the best of our knowledge, no prior study reported such an improvement

- It is generic enough to be extended to other similar data sets

- This approach can enable organizations to follow the encouragements stated in the NIH data sharing policy