Cluster-based Anonymization of Knowledge Graphs

Presenter: Anh-Tu Hoang (ahoang@uninsubria.it)

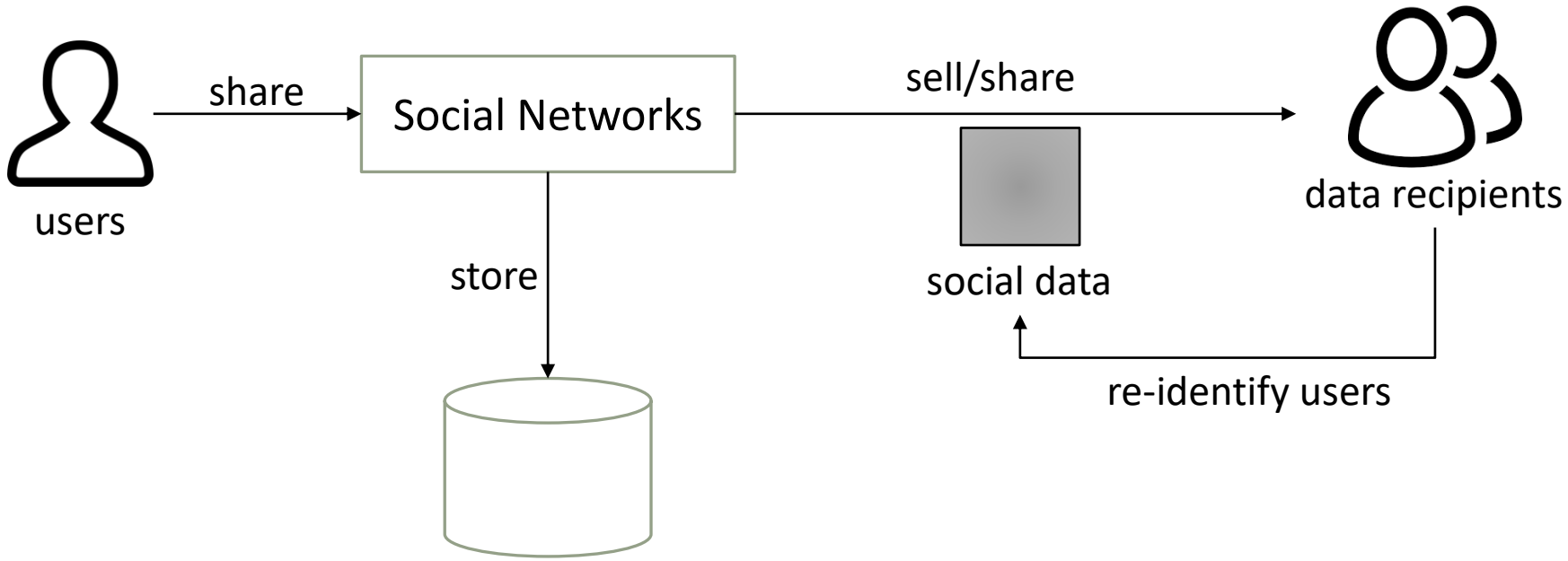
Anh-Tu Hoang, Barbara Carminati, Elena Ferrari

DiSTA, University of Insubria, Italy

Agenda

- Introduction
- Related works
- The k-Attribute Degree
- Information Loss Metrics
- Cluster-based Anonymization Algorithm (CKGA)
- Experiments
- Conclusion

Risks of sharing social data



What type of data can be anonymized?

attributes' values

User	Age	Job
Ken	18	Student
Mary	19	Student
Henry	55	Professor
Tom	40	Professor

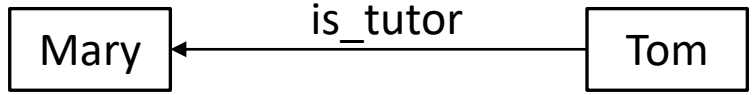
relational data

protect

k-anonymity [1]

follow relationships

is_tutor relationships



directed graph

directed graph

protect

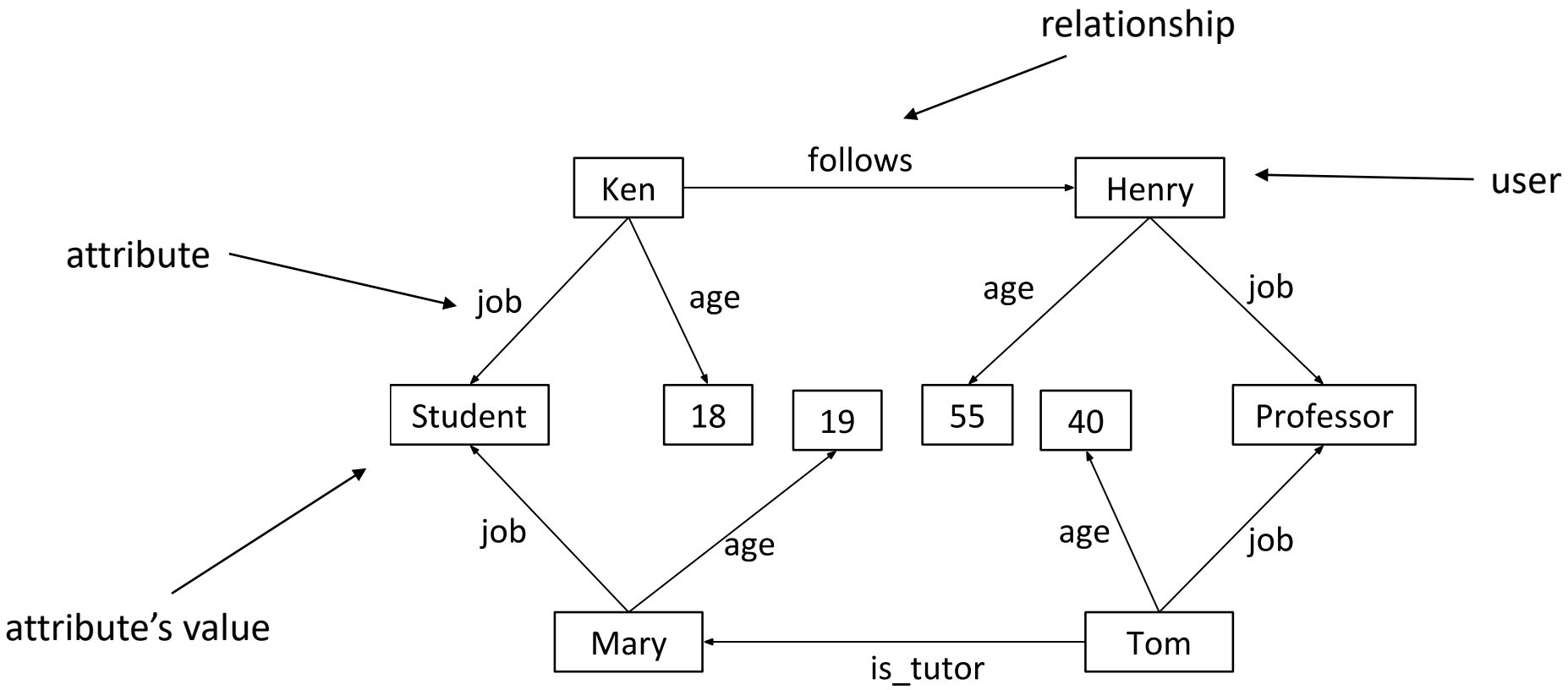
protect

Paired k-degree [2]

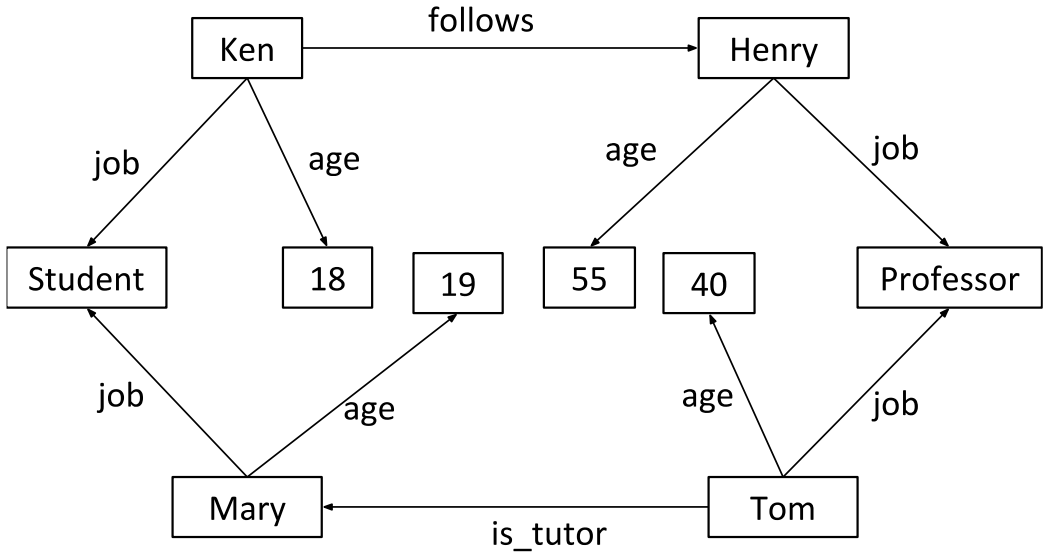
k-In&Out-Degree Anonymity [3]

anonymization solutions for both users' attributes and many types of relationships are missing.

Knowledge Graphs (KG)

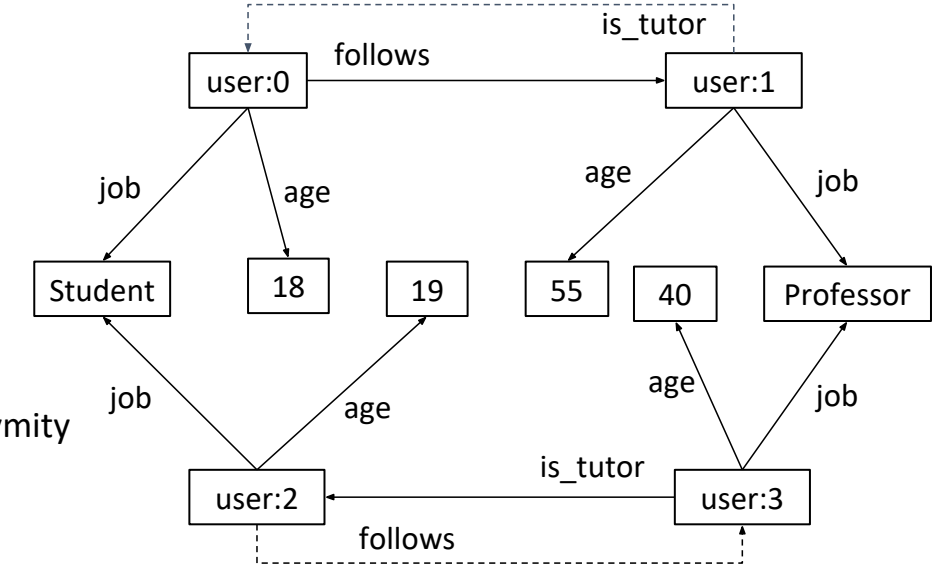


How to anonymize knowledge graphs?



anonymize

Paired k-degree
K-In&Out-Degree Anonymity



user:0 and user:2 have the same out-/in-degrees for: job, age, follows, is_tutor

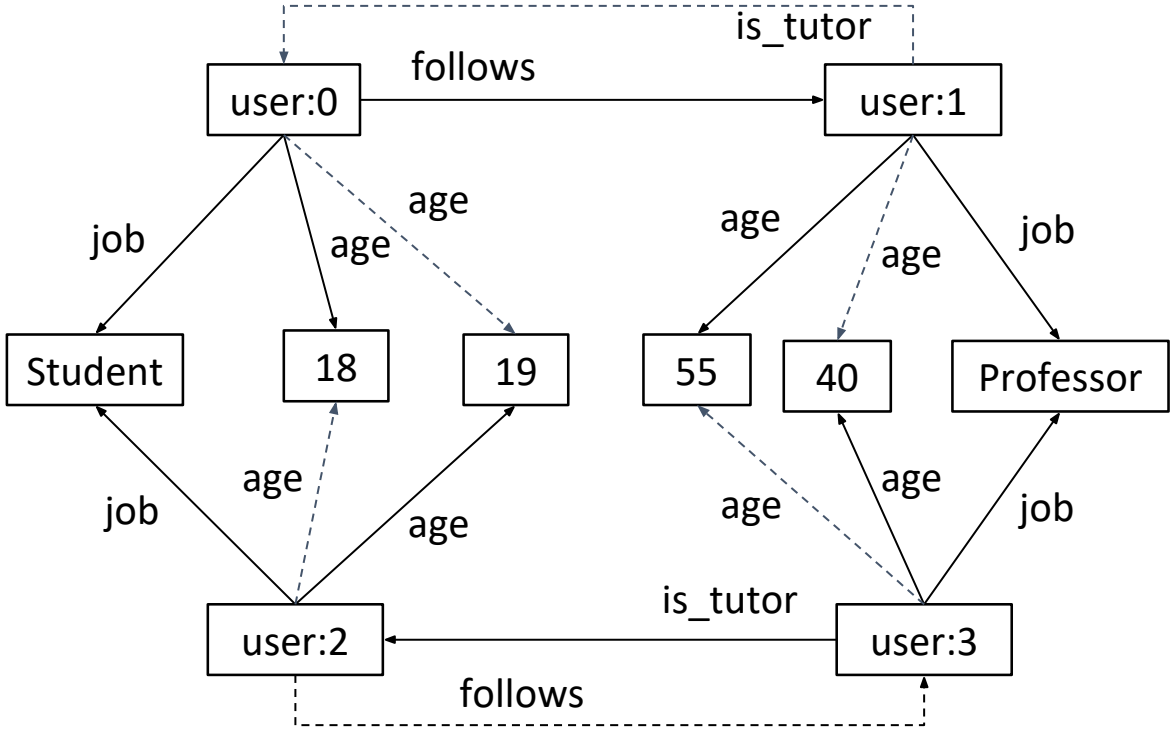
if adversaries know Ken's age and the number of users he follows, they can re-identify user:0 is Ken.

Our contributions

- k-AttributeDegree (k-ad): protect users' identities in anonymized KG.
- Two information loss metrics:
  - Attribute & Degree Information Loss Metric.
  - Attribute Truthfulness Information Loss Metric.
- **C**lusters-Based **K**nowledge **G**raph **A**nonymization Algorithm (CKGA).
- We prove that our algorithm always generate anonymized KGs satisfying k-ad.

k-Attribute Degree (k-ad)

k-ad ensures that attributes' values and relationships' out-/in-degrees of users are indistinguishable from those of k-1 other users.



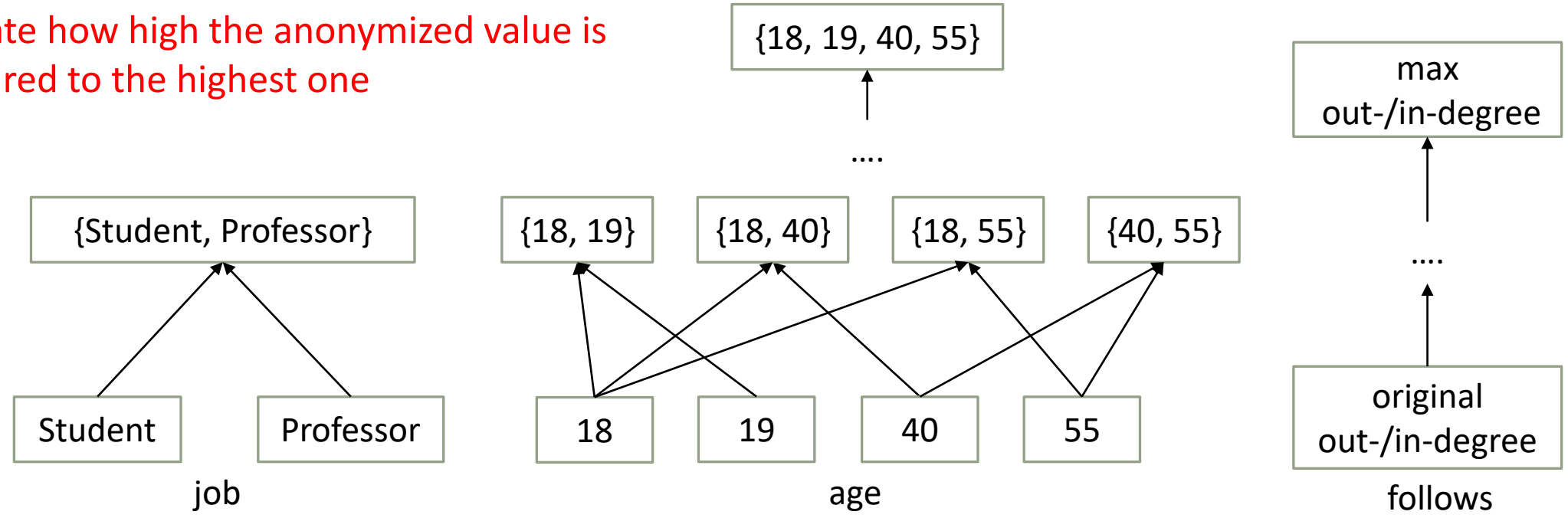
k=2: attributes' values and relationships' out-/in-degrees of user:0 and user:2 are identical

user:1 and user:3

Attribute & Degree Information Loss

what if a Professor has age 18 after anonymization?

calculate how high the anonymized value is compared to the highest one



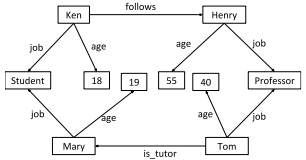
categorical attribute

numerical attribute

relationships

Attribute Truthfulness Information Loss

calculate the percentage of truthful associations of a user's attributes



original knowledge graph

train (PyTorch)

truthful indicator

(age, 19)

(age, 18)

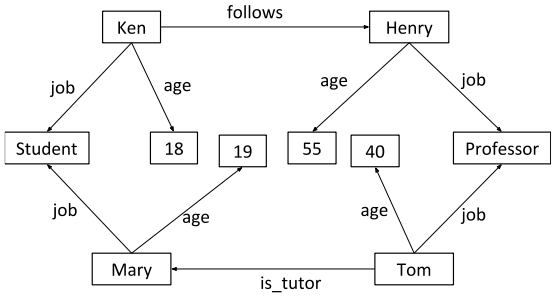
(job, Professor)

(job, Student)

how truthful a 19-year-old Student is

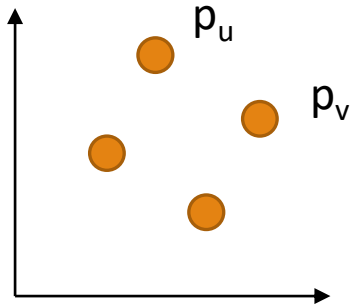
1: (age, 19), (job, Student) is truthful
0: (age, 19), (job, Student) is untruthful

Clusters-Based Anonymization (1)



original knowledge graph

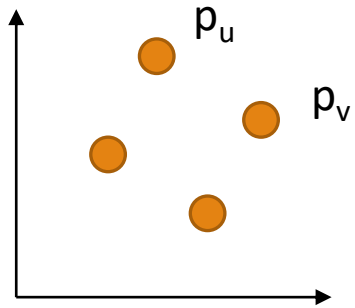
train



Euclidean distance(p_u, p_v) ~ InfoLoss(u, v)

Information loss of making attributes' values and relationships' out-/in-degrees of u and v identical

Clusters-Based Anonymization (2)



generated points

Clusters Generation

clustering

k-Means Partition

k-means
HDBSCAN

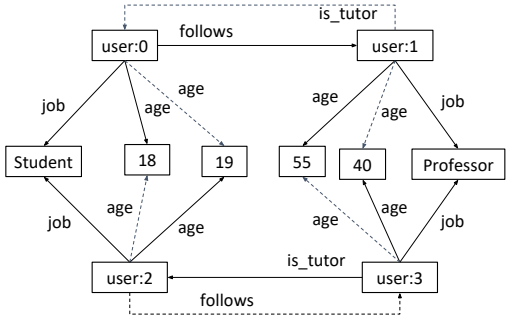
ensure clusters have from k to 2*k-1 users

Knowledge Graph Generalization

Attributes Generalization

Relationships Generalization

add/remove edges to make attributes' values and relationships' out-/in-degrees of users in the same cluster identical



anonymized knowledge graph

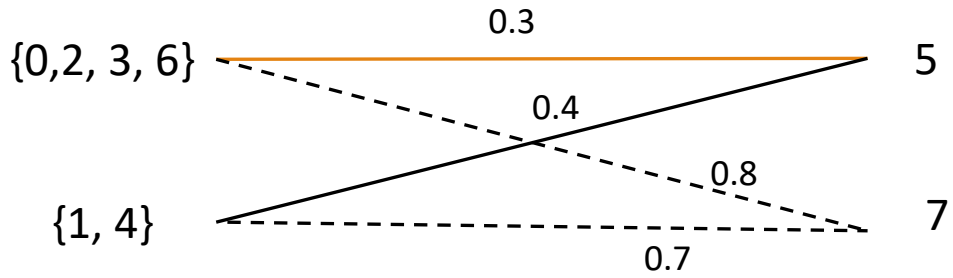
k-Means Partition (KP)

split clusters that have less than k users

k=2 clusters: {0, 2, 3, 6} {1, 4} {5} {7}

generated from a clustering algorithm (e.g., k-means, HDBSCAN)

assign new clusters



distance <= max_dist

distance > max_dist

split clusters that have at least 2*k users

{0,2,3,6,5}
{1,4}

balanced k-means

{0,3,6}
{2,5}
{1,4}

Knowledge Graph Generalization

generalize
cluster {2,5}

original edges

(2,age,20)
(2,job,Student)
(5,age,22}
(5,job,Engineer)

add
attribute edges

anonymized edges

(2,age,20)
(2,age,22)
(2,job,Student)
(2,job,Enginner)
(5,age,20)
(5,age,22)
(5,job,Student)
(5,job,Engineer)

2's age is either 20 or 22
2's job is either Engineer or Student

attribute edges

(2,follows,6)
(2,follows,5)
(5,follows,1)
(1,follows,5)

add/remove
relationship edges

(2,follows,6)
(3,follows,2)
~~(2,follows,5)~~
(5,follows,1)
(1,follows,5)

minimize the number of added/removed relationship edges

relationship edges

Evaluation

- 5 real-life data sets: Email-Eu-core[6], Google+[6], Freebase[7], Email-temp[6], DBLP[6].
- Tune parameters.
- Evaluate the truthfulness of KGs.
- Compare to CDGA[4], DGA[2].

How good are the generated vectors?

The higher the number of dimensions, the fewer differences between Euclidean distances and information loss.

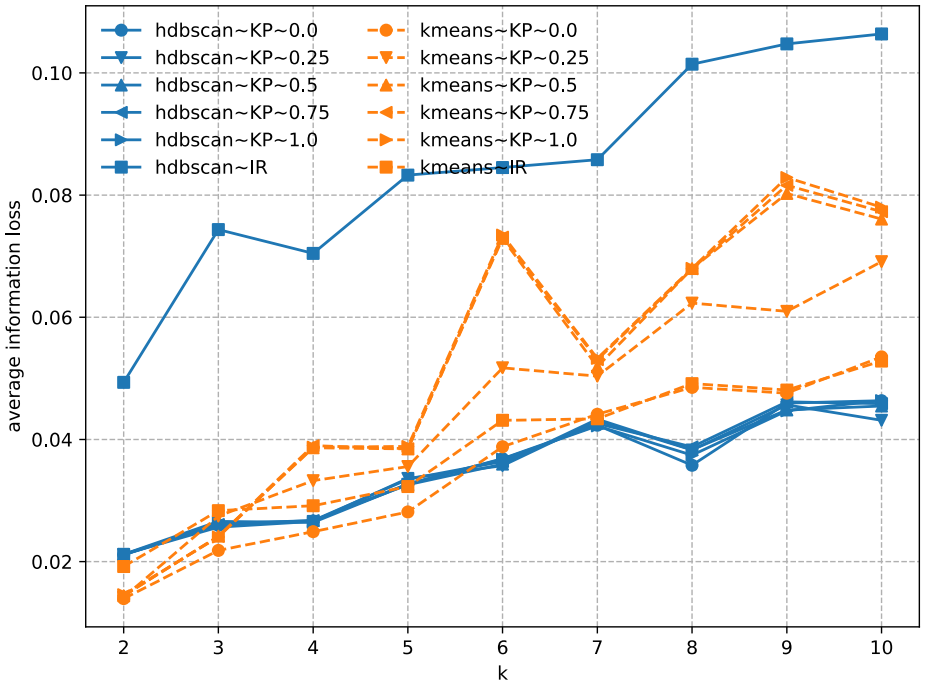
Data set	$d_2 = 2$	$d_2 = 10$	$d_2 = 50$
Email-Eu-core	0.0046 ($\pm$0.0038)	0.0012 ($\pm$0.0015)	**0.0005** ($\pm$**0.0009**)
Google+	0.0099 ($\pm$0.0083)	0.0054 ($\pm$0.0040)	**0.0008** ($\pm$**0.0012**)
Freebase	0.0072 ($\pm$0.0073)	0.0036 ($\pm$0.0032)	**0.0003** ($\pm$**0.0010**)
Email-temp	0.0030 ($\pm$0.0030)	0.0019 ($\pm$0.0012)	**0.0001** ($\pm$**0.0002**)
DBLP	0.0073 ($\pm$0.0021)	0.0031 ($\pm$0.0011)	**0.0002** ($\pm$**0.0001**)

d_2: the number of dimensions of generated vectors

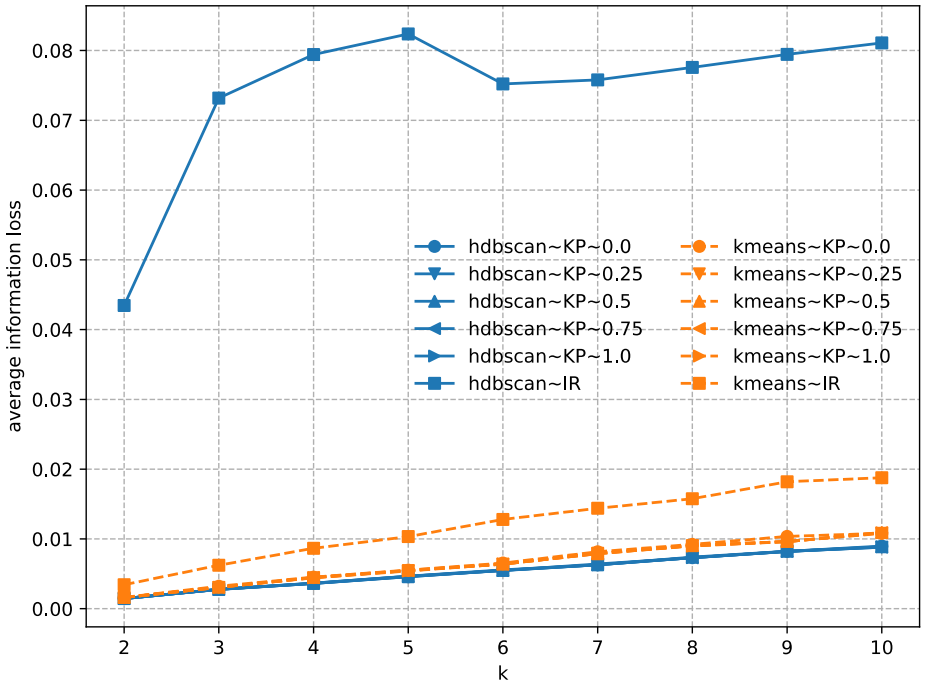
The generated data points are good enough to be used in clustering algorithms.

What clustering algorithm is good for anonymizing KGs?

- clusters that have more than 2*k users results in high information loss.
- decreasing the maximum distance decreases information loss since it removes outliers.
- k-means generates better quality clusters as these clusters are smaller than those generated from HDBSCAN.



Email-Eu-core

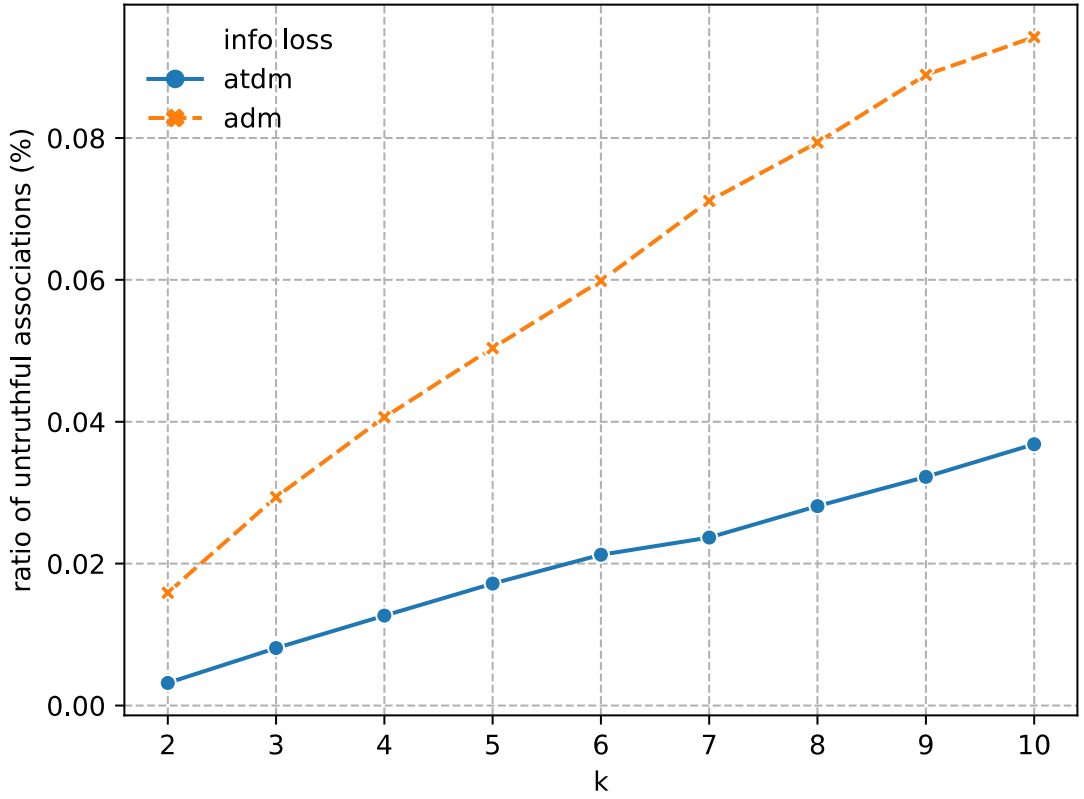


Freebase

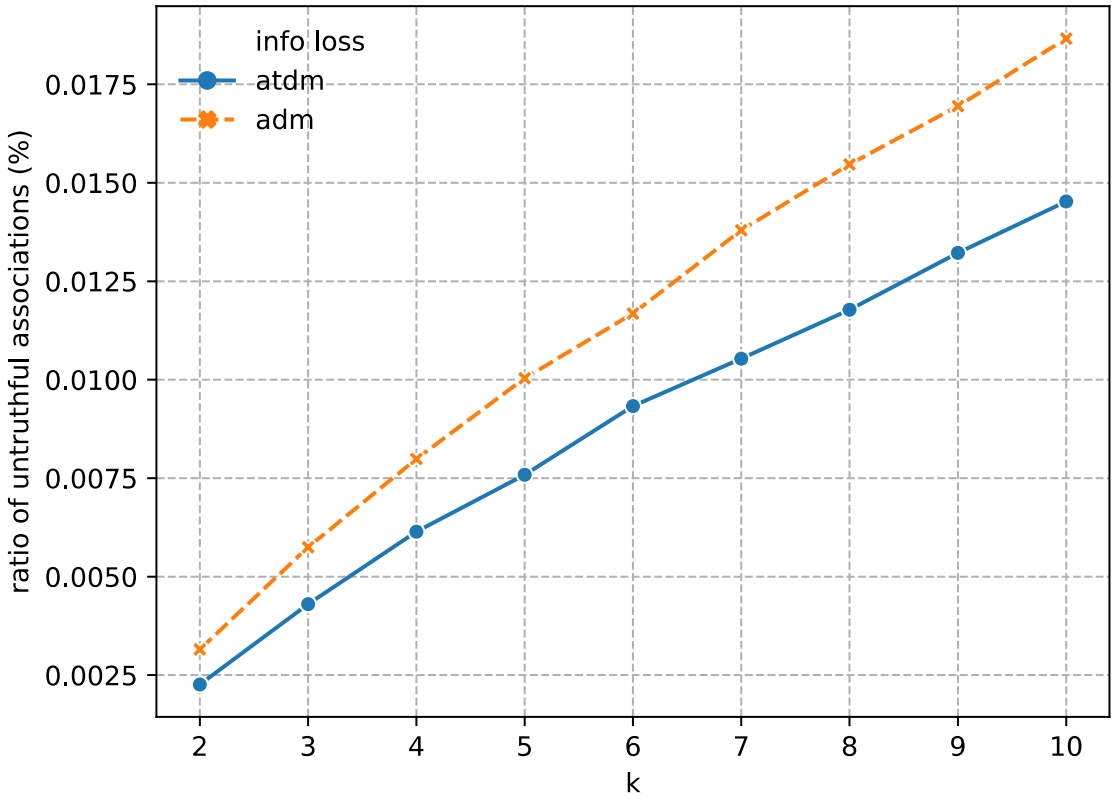
IR: removes clusters that have less than k users, KP: k-Means Partition Algorithm

How effective is ATDM comparing to ADM?

Our Attribute Truthfulness Information Loss Metric is effective enough to increase the truthfulness of users' attributes.



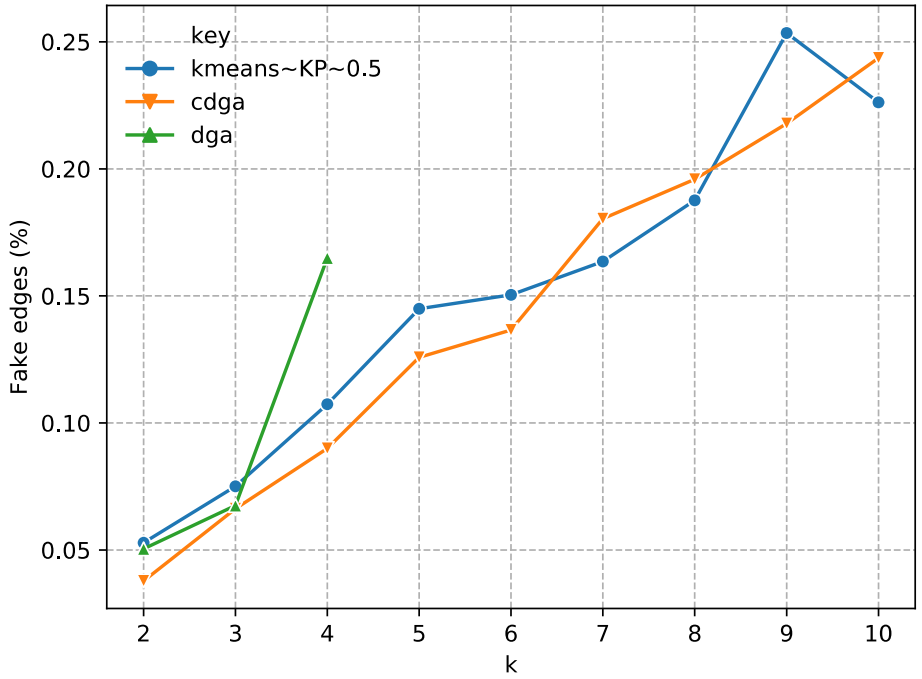
Google+



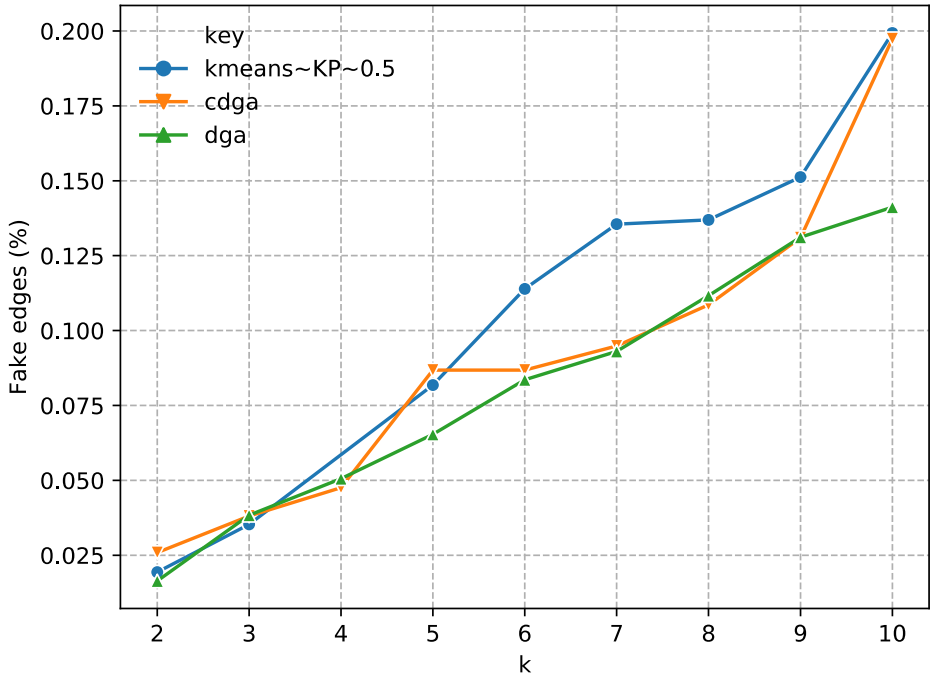
Freebase

Comparison with the state-of-the-art

Anonymized directed graphs generated by our algorithm contain the similar number of fake edges comparing to that of those generated by DGA and CDGA.



Email-temp



DBLP

Performance of clustering algorithms of CKGA and CDGA

- Using generated points increases performance of clusters generation algorithm.

	Email-temp		DBLP	
k	CDGA	CKGA	CDGA	CKGA
2	680.8	6.5	91,855.4	607.9
3	825.7	4.4	111,181.6	420.4
4	893.3	3.2	118,269.7	629.9
5	929.7	2.7	122,246.1	581.5
6	950.4	2.3	124,145.7	540.3
7	977.4	1.9	126,542.7	483.3
8	987.9	1.8	127,343.4	474.9
9	1,000.7	1.5	128,254.3	457.6
10	1,007.1	1.5	128,727.9	398.9

generated points for Email-temp: 408s and DBLP: 40,415s.

Conclusion

- k-Attribute Degree:
  - protect users' identities when adversaries know attributes' values and out-/in-degrees of their victims.
- Information Loss Metrics: ADM, ATDM.
- The Cluster-based Anonymization Algorithm for Knowledge Graphs (CKGA).
  - Data providers can use any clustering algorithm to generate anonymized KGs.
  - Can replace anonymization solutions for relational data and directed graph.
- Future works:
  - l-diversity[5] solution for knowledge graphs.
  - Sequentially publishing of knowledge graphs.

References

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570.

[2] Casas-Roma, Jordi, et al. "k-Degree anonymity on directed networks." *Knowledge and Information Systems* 61.3 (2019): 1743-1768.

[3] Zhang, Xiaolin, et al. "Large-scale dynamic social network directed graph k-in&out-degree anonymity algorithm for protecting community structure." *IEEE Access* 7 (2019): 108371-108383.

[4] Hoang, Anh-Tu, Barbara Carminati, and Elena Ferrari. "Cluster-Based Anonymization of Directed Graphs." *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2019.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *22nd International Con- ference on Data Engineering (ICDE'06)*, April 2006, pp. 24–24.

Thank you for your attention