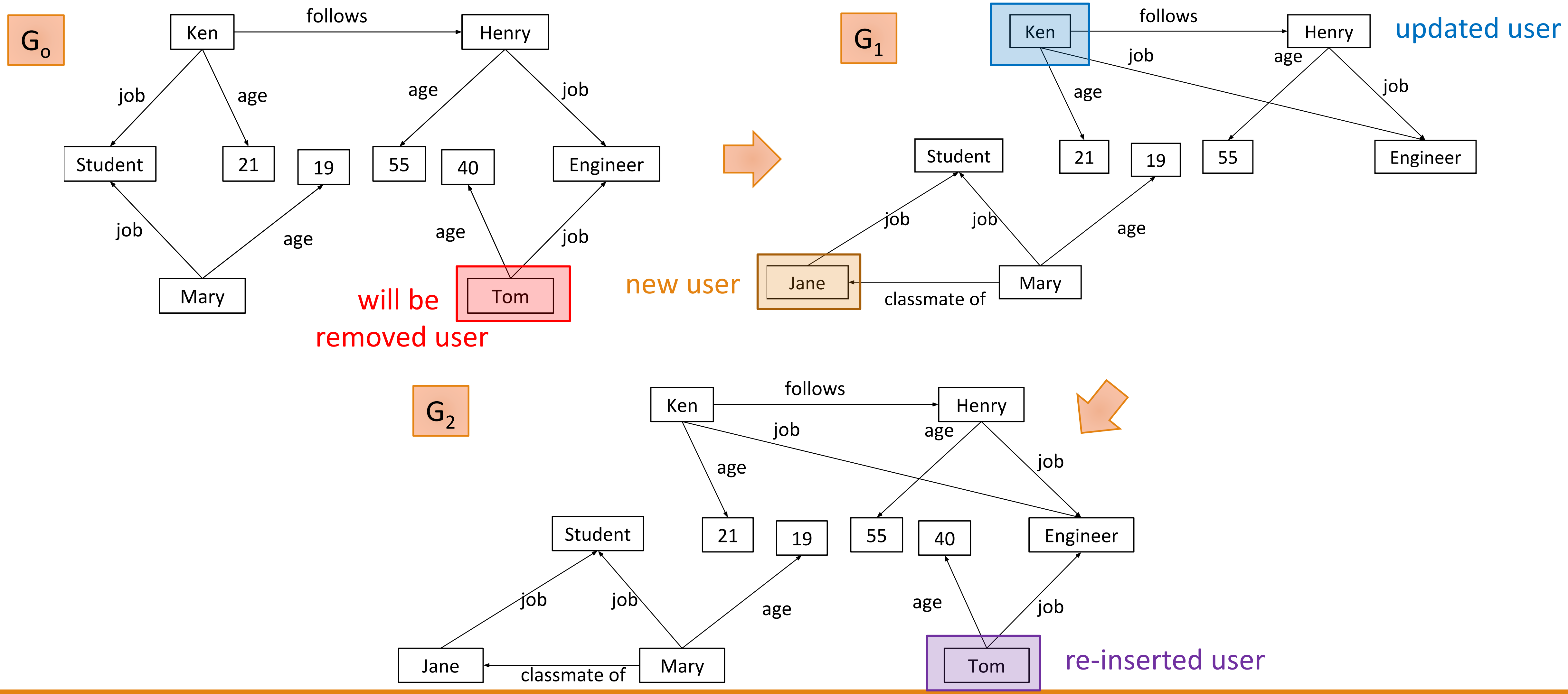
Privacy-Preserving Sequential Publishing of Knowledge Graphs

Anh-Tu Hoang, Barbara Carminati, Elena Ferrari

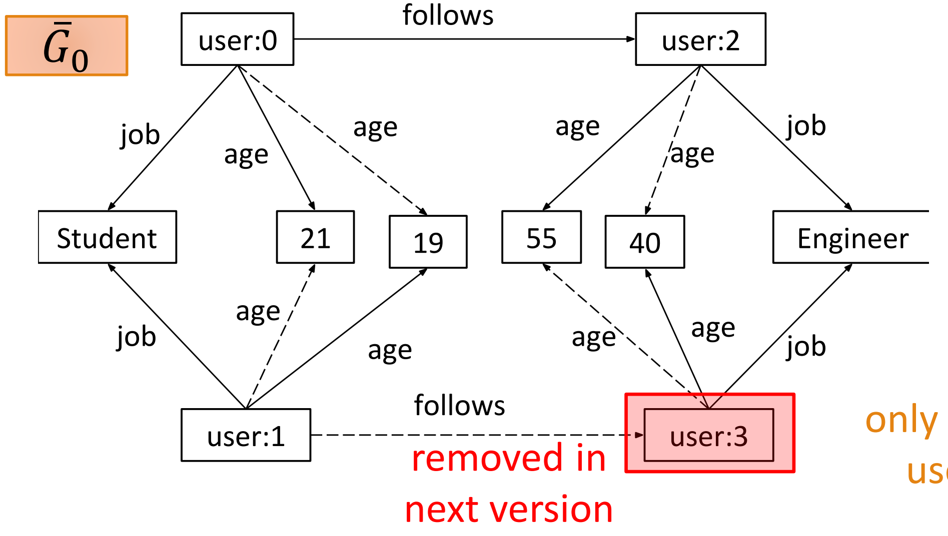
{ahoang, barbara.carminati, elena.ferrari}@uninsubria.it

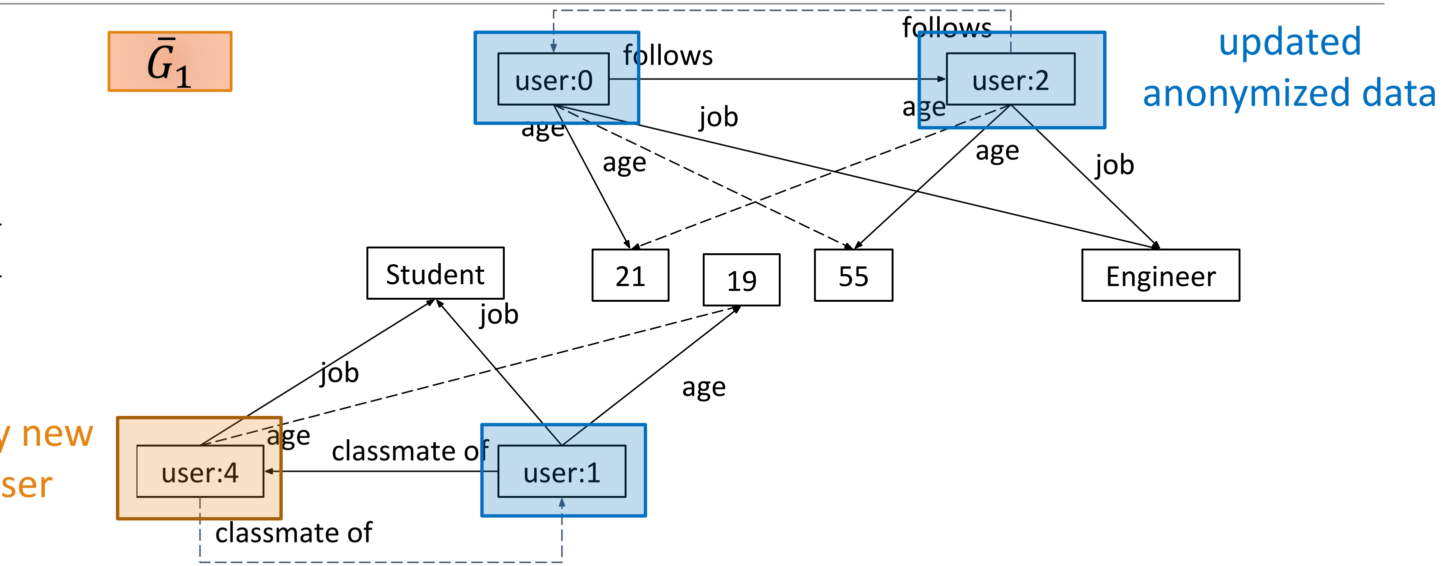
DiSTA, University of Insubria, Italy

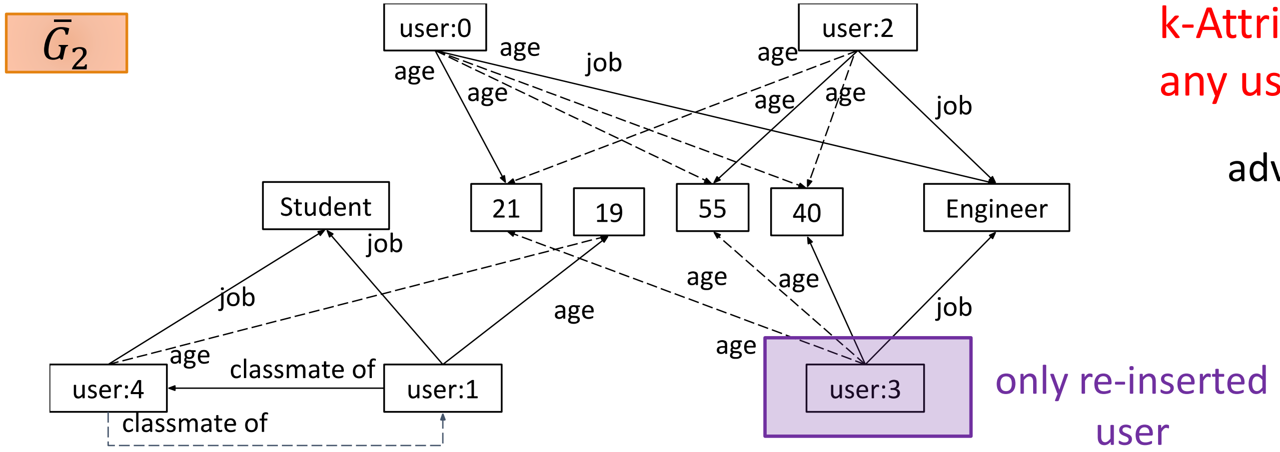
How to sequentially anonymize knowledge graphs?



Anonymizing independently does not work







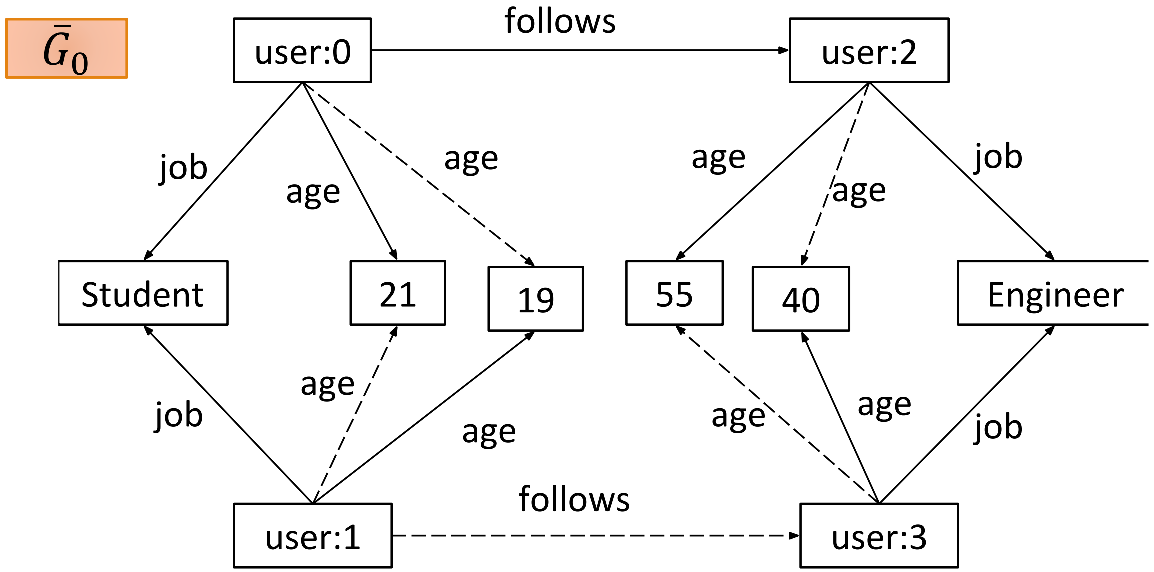
k-Attribute-Degree[1]: attribute values and relationship degrees of any user are indistinguishable from that of k-1 other users.

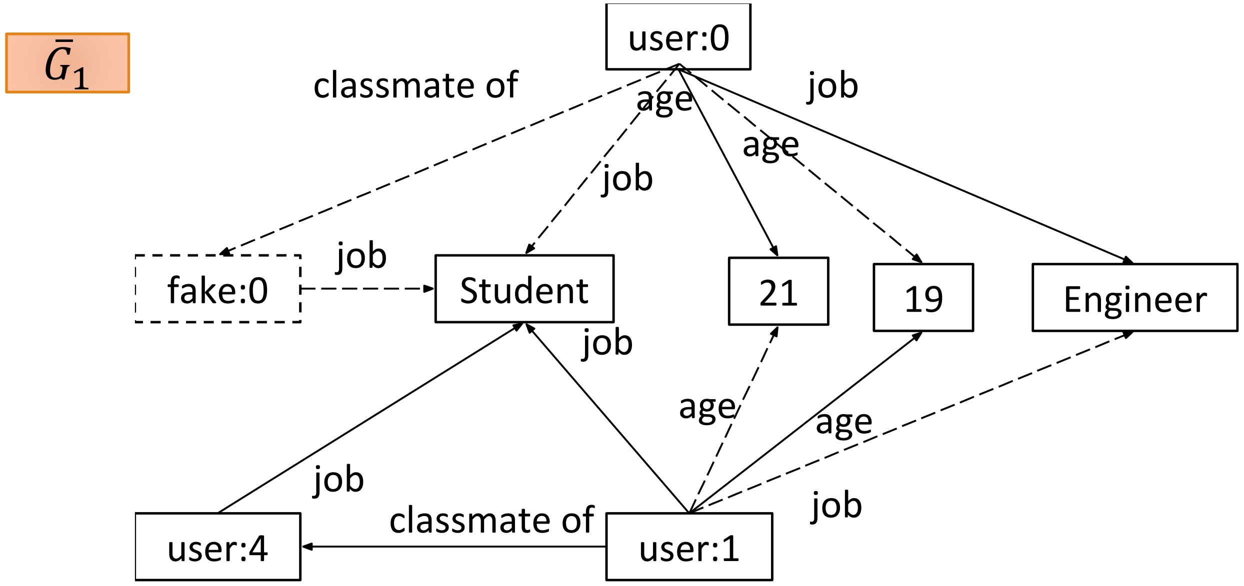
adversaries can re-identify: user:0 user:2 user:1 user:3 user:4

[1] Anh-Tu Hoang, Barbara Carminati, and Elena Ferrari. "Clusters-Based Anonymization of Knowledge Graphs". Proceedings of the International Conference on Applied Cryptography and Network Security (ACNS), Italy, 2020.

Time-Varying k-Attribute Degree

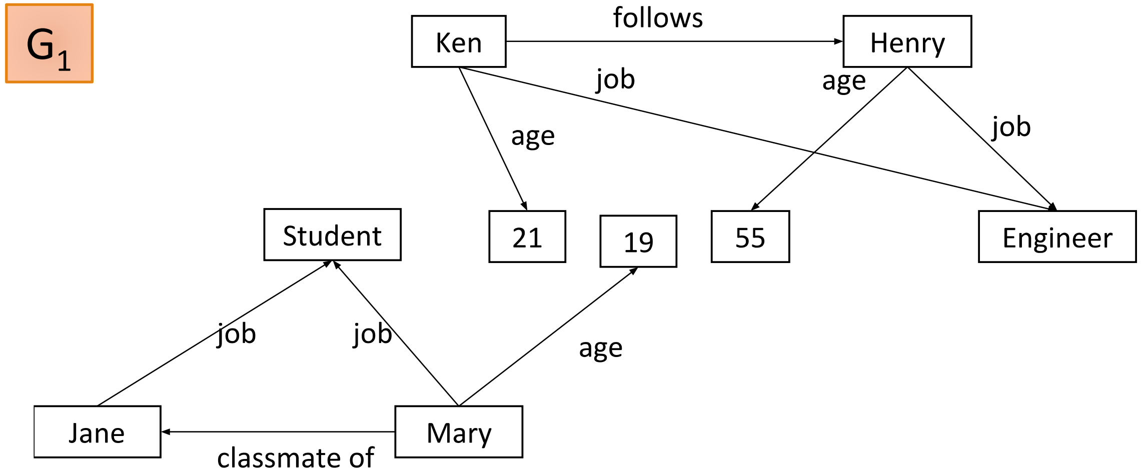
ensure that for every **user in w continous anonymized KGs**, the changes of his/her attributes' values and degrees are identical to those of k-1 other users in these KGs.





Time-Varying Anonymization - CTKGA (1)

G_1

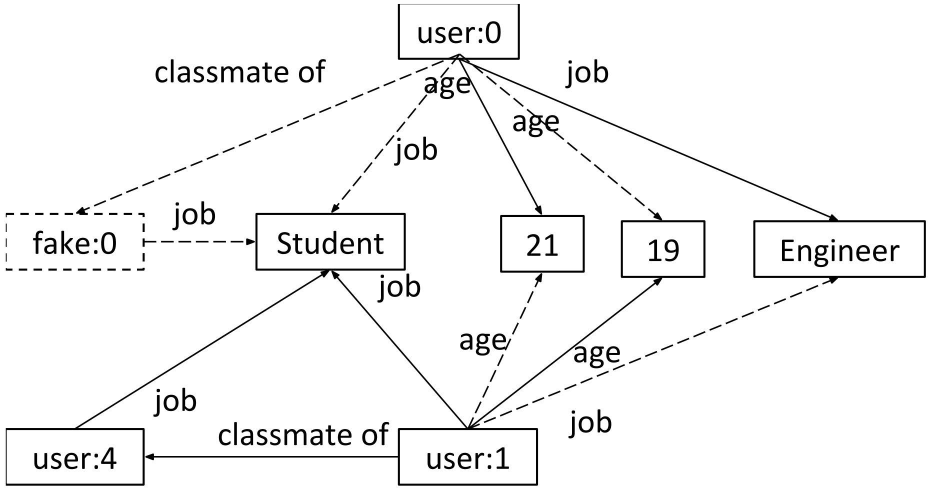


clusters generation

{user:0, user:1}
{user:4, fake:0}

clusters

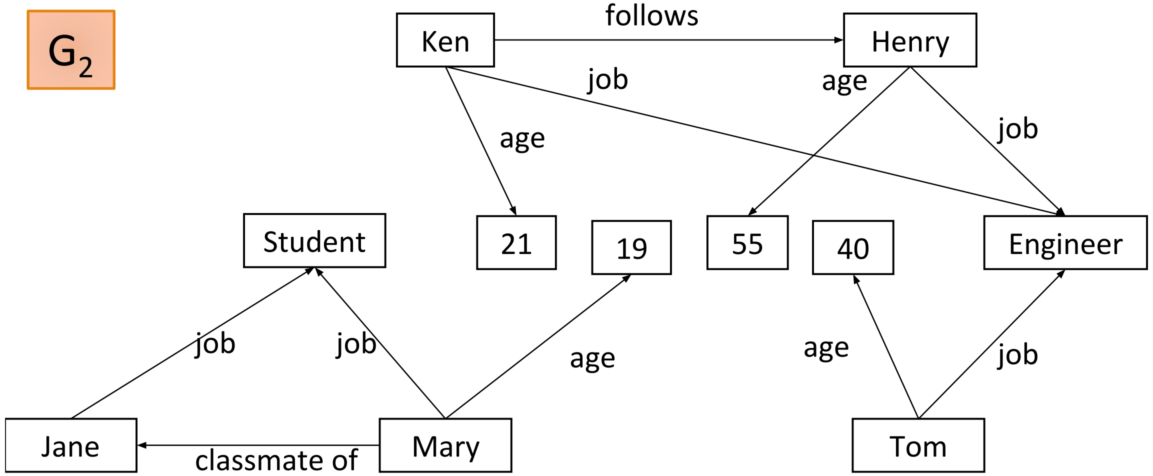
knowledge graph generalization



Info	**Users**
(I_0^0)	{user:0, user:1}
(I_0^1)	{user:2, user:3}

ADS-Table H_0^2

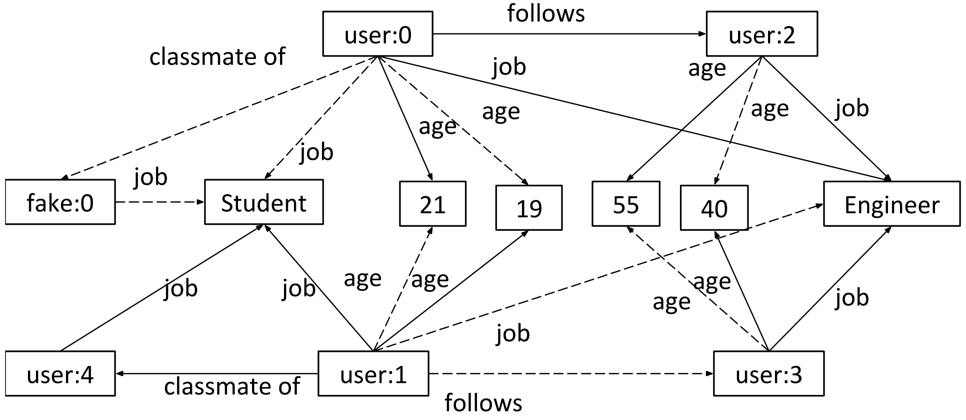
Time-Varying Anonymization - CTKGA (2)



clusters generation

{user:0, user:1}
{user:4, fake:0}
{user:2, user:3}
clusters

knowledge graph generalization

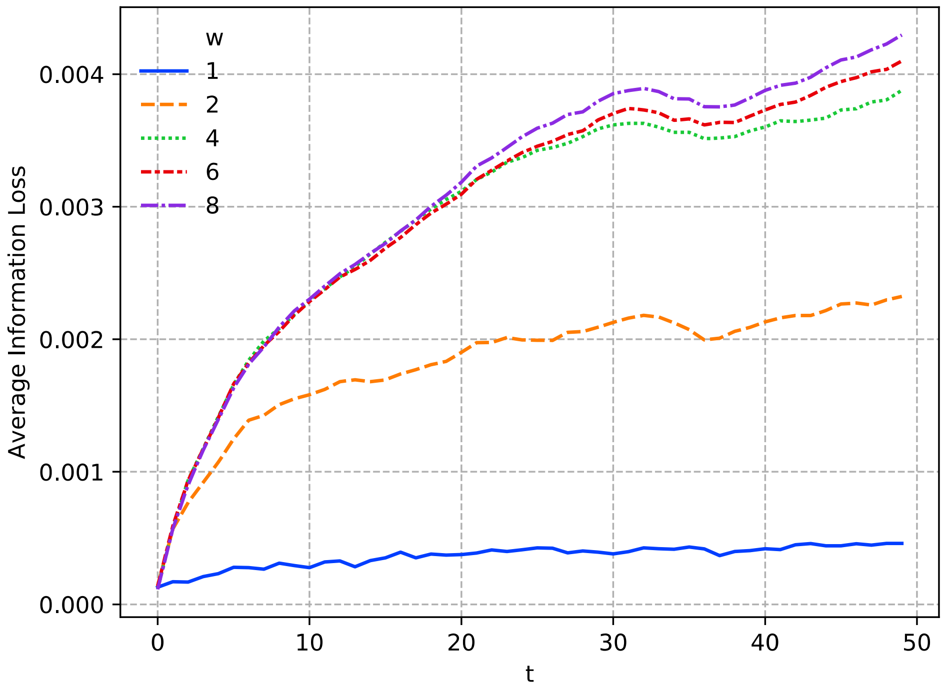


Info	Users
(I_0^0, I_1^0)	{user:0, user:1}
$(I_0^1, \emptyset)$	{user:2, user:3}
$(\emptyset, I_1^1)$	{user:4, fake:0}

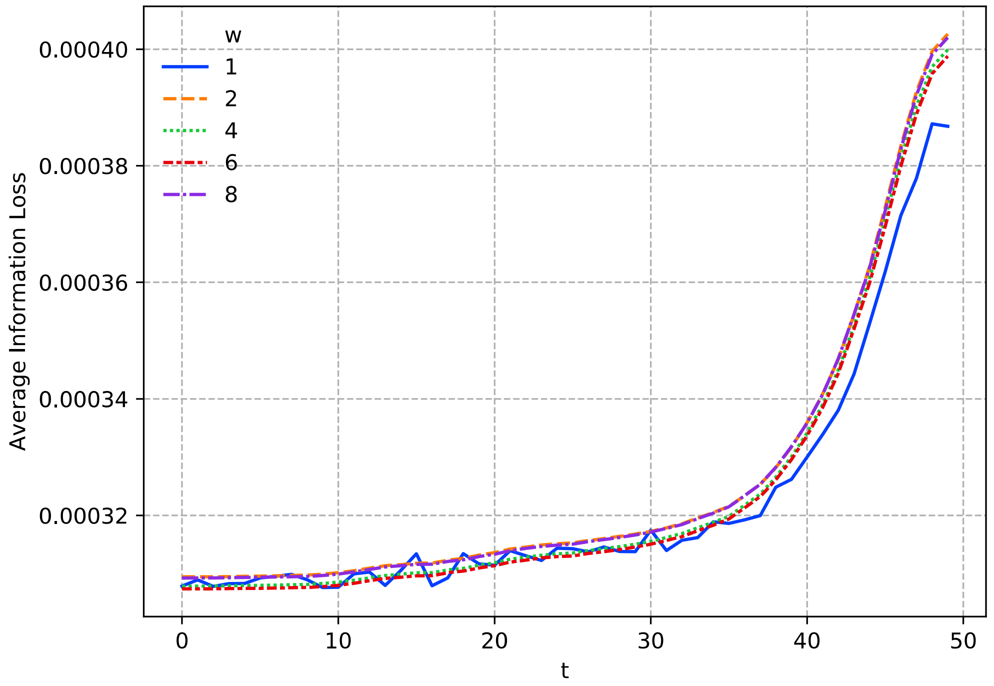
ADS-Table H_1^2

Impact of ω

- ❖ increasing ω decreases the quality of anonymized KGs.
- ❖ when ω is high enough, the quality of anonymized KGs does not decrease too much while users are protected with higher constraints.



Email-temp[2]



Yago[3]

[2] A. Paranjape, A. et al. “Motifs in temporal networks”. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.
[3] García-Durán, Alberto, et al. "Learning sequence encoders for temporal knowledge graph completion”. *arXiv preprint arXiv:1809.03202,* 2018.

Conclusion & Future work

❖ We presented attacks and defenses for sequential anonymizing knowledge graphs.

❖ Our solutions are flexible enough to allow data providers to insert/remove/update/re-insert user information.

❖ Future work:

- Protecting users’ sensitive values (i.e., disease, salary) in knowledge graphs.
- Applying differential privacy to design safe machine learning algorithms for knowledge graphs.
- Allowing users to specify their own k values.

References

[1] Anh-Tu Hoang, Barbara Carminati, and Elena Ferrari. "Clusters-Based Anonymization of Knowledge Graphs". Proceedings of the International Conference on Applied Cryptography and Network Security (ACNS), Italy, 2020.

[2] A. Paranjape, A. et al. “Motifs in temporal networks”. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.

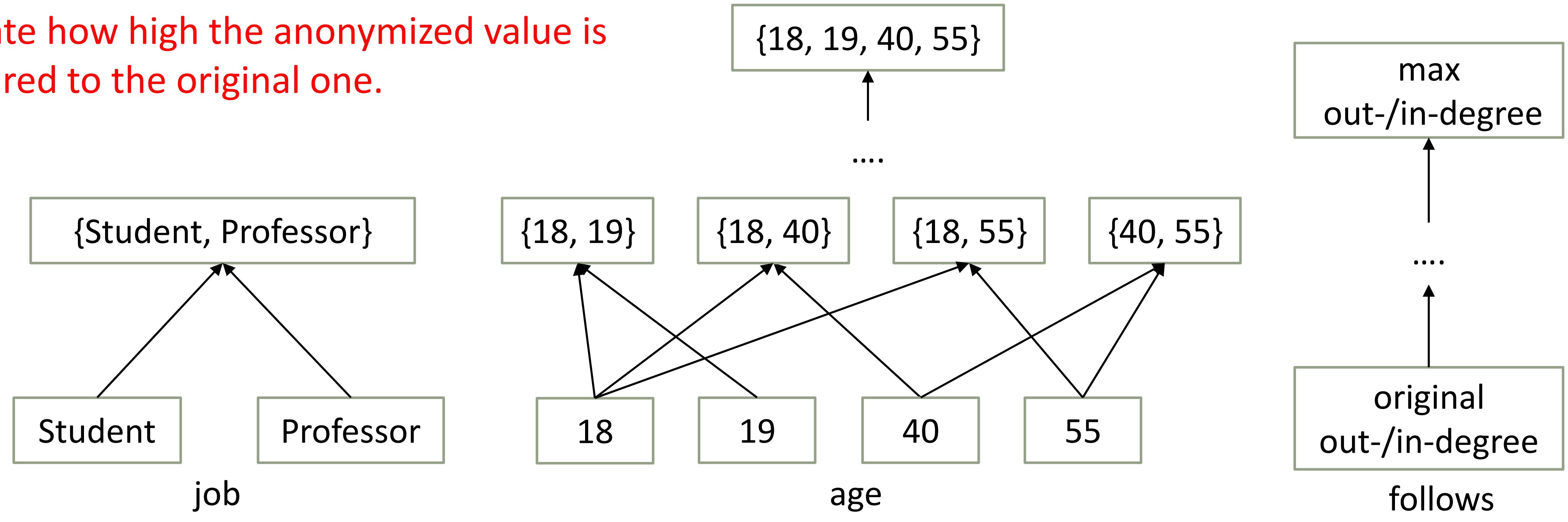
[3] García-Durán, Alberto, et al. "Learning sequence encoders for temporal knowledge graph completion”. *arXiv preprint arXiv:1809.03202,* 2018.

Thank you for your attention

Attribute & Degree Information Loss (ADM)

what if a Professor has age 18 after anonymization?

calculate how high the anonymized value is compared to the original one.



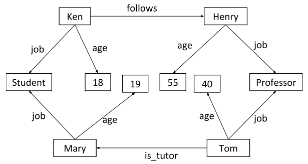
categorical attribute

numerical attribute

relationships

Attribute Truthfulness Information Loss (ATDM)

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



original
knowledge graph

train (PyTorch)

truthfulness
indicator

(age, 19)

(age, 18)

(job, Student)

(job, Professor)

how truthful is a 19-year-old Professor?

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

Out- and In-Degree Information Loss (DM)

Out-degree information loss of a user *u*

$$DM'^{\overline{G}}_o(u) = \frac{|I^{\overline{G}}_o(u) - I^{G}_o(u)|}{|V|}$$

Out-degree information loss if we make the out-degree of two users u, v identical

$$DM^{\overline{G}}_o(u,v) = \frac{DM'^{\overline{G}}_o(u) + DM'^{\overline{G}}_o(v)}{2}$$

$$I^{\overline{G}}_o(u) = I^{\overline{G}}_o(v)$$

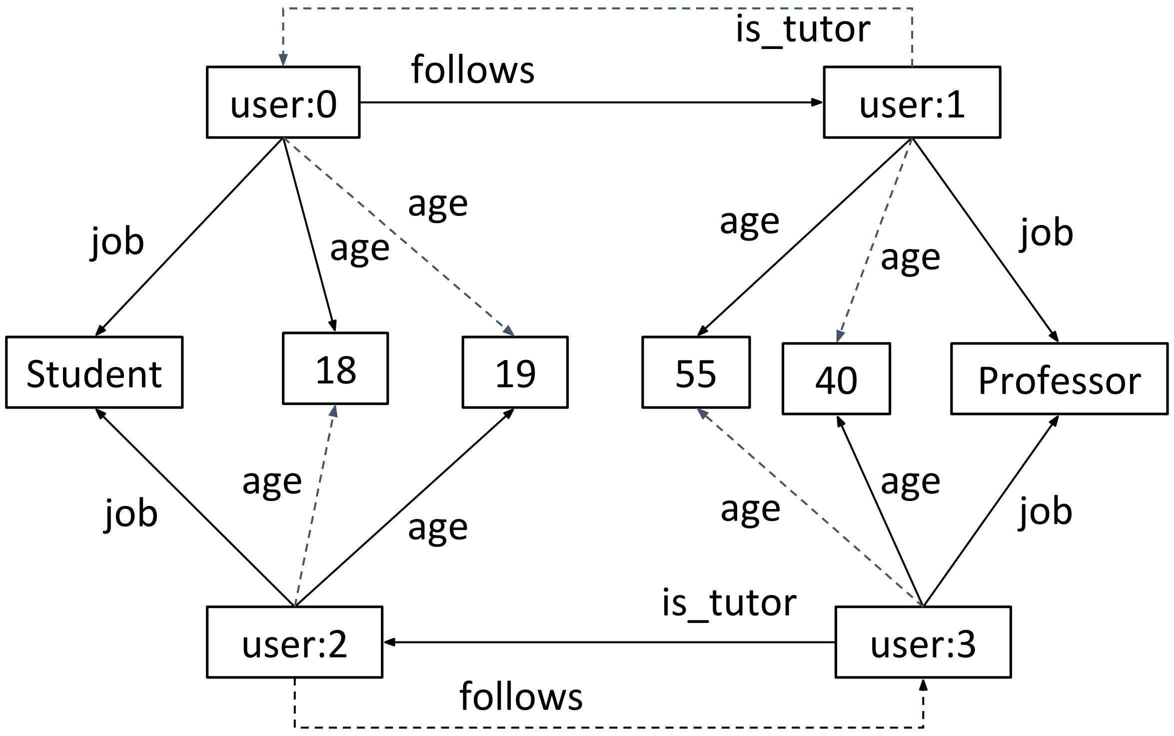
Combine Out- and In-Degree information loss of making out- and in-degree of two users u, v identical

$$DM^{\overline{G}}(u,v) = \alpha \times DM^{\overline{G}}_o(u,v) + (1-\alpha) \times DM^{\overline{G}}_i(u,v)$$

similar to DM_o

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

k-Means Partition (KP)

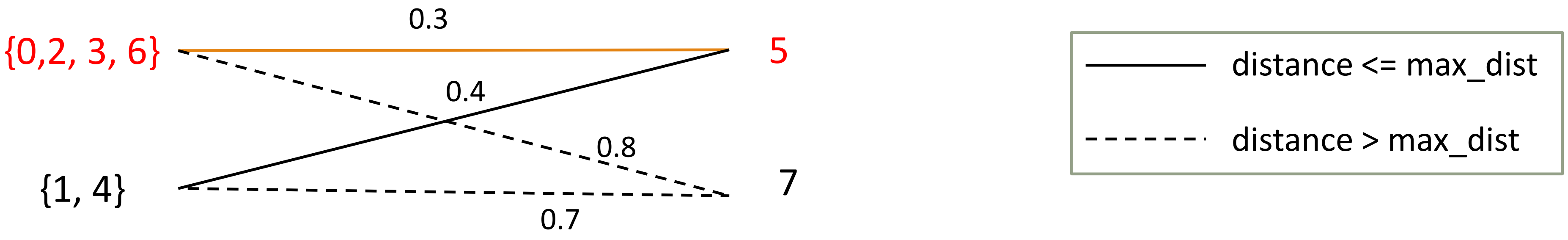
remove clusters that have less than k users

k=2

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

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

assign new clusters



split clusters that have at least 2*k users

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

balanced k-means[7]

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

[7] Malinen, Mikko I., et al. "Balanced k-means for clustering”. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2014.

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

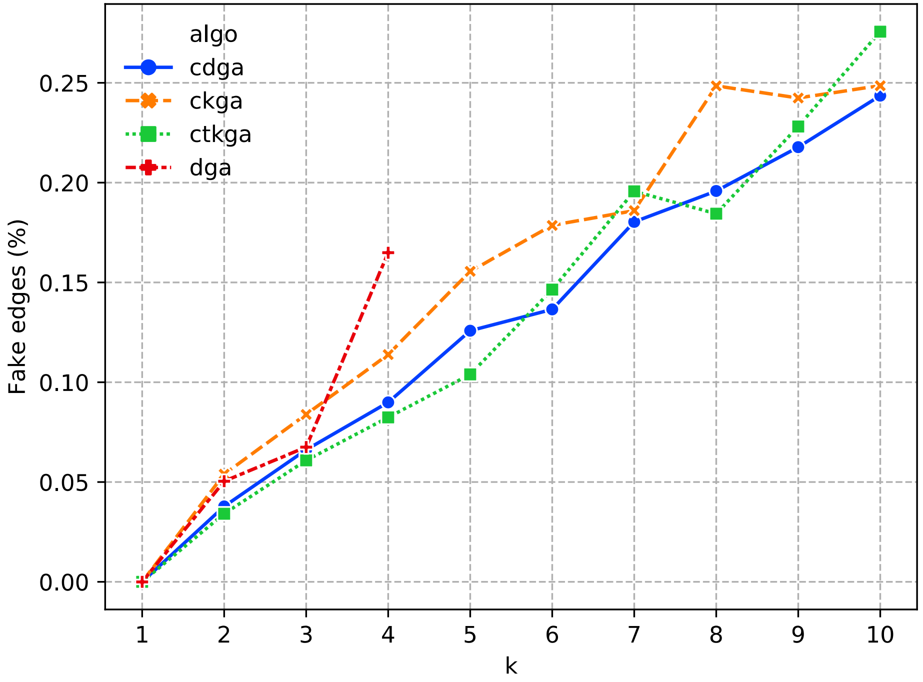
(2,follows,6)
(3,follows,2)
~~(2,follows,5)~~
(5,follows,1)
(1,follows,5)

relationship edges

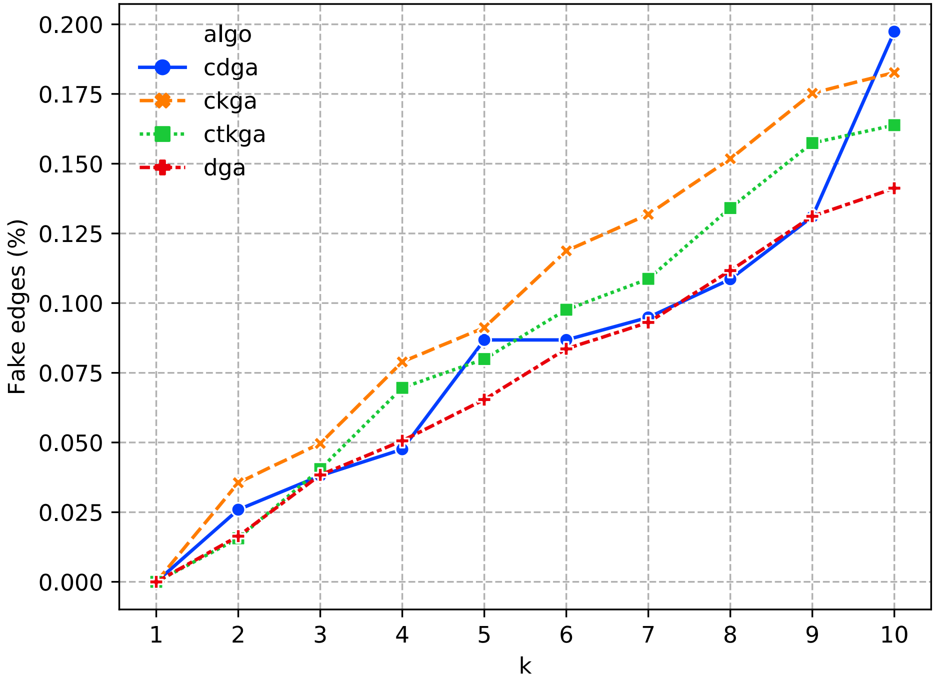
minimize the number of added/removed relationship edges

Compare to CKGA, CDGA, DGA [2]

❖CTKGA adds similar the number of fake edges to that of CKGA.

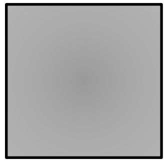


Email-temp

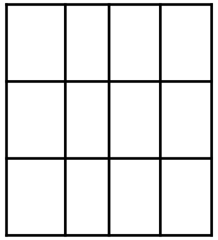


DBLP

Time-Varying Clusters-Based Anonymization



distance matrix



ADS-Table

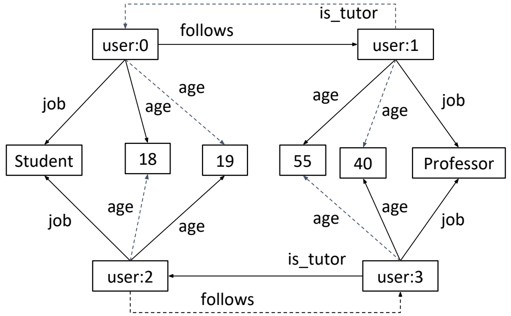
Clusters Generation

clustering

k-Medoids Partition

k-medoids
HDBSCAN

Knowledge Graph Generalization



anonymized knowledge graph

- ensure clusters have from k to 2*k-1 users.
- users in the same cluster have the same attributes and degrees in w-1 previous KGs.

add/remove edges such that all users in the same clusters have the same values for attributes and out-/in-degrees.

store users' attributes and degrees in w-1 previous KGs.