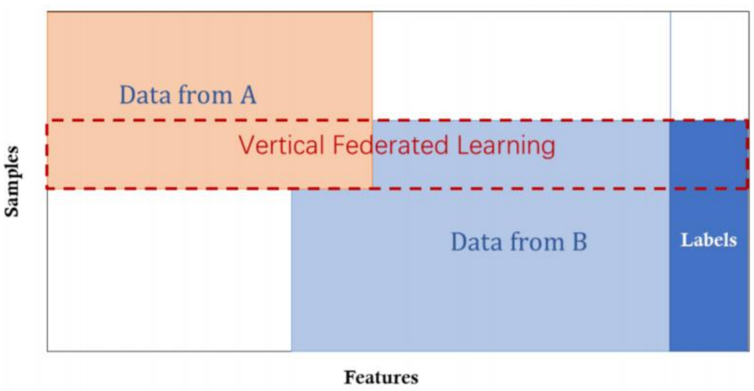
A Coupled Design of Exploiting Record Similarity for Vertical Federated Learning

Zhaomin Wu, Qinbin Li, Bingsheng He
National University of Singapore

Vertical Federated Learning (VFL)



(Yang et al. TIST 2019)

Share the same sample space
Own a subset of features
Only one party has labels

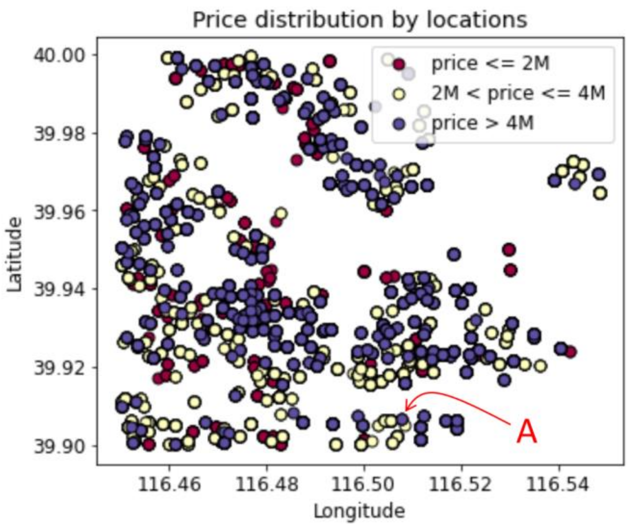
How to determine which instances should be involved in training?
Privacy-Preserving Record Linkage (PPRL) [1]

How existing studies use PPRL in VFL?
Train exactly/top1 matched records

According to the study in German record linkage center [2], 72.7% of the applications suffer information loss by exact/top1 linkage

[1] Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. Handbook of big data technologies, 851-895.
[2] Manfred Antoni and Rainer Schnell. The past, present and future of the german record linkage center (grlc). Jahrbücher für Nationalökonomie und Statistik, 239(2):319–331, 2019.

Record Linkage



Housing price by geolocations in Beijing

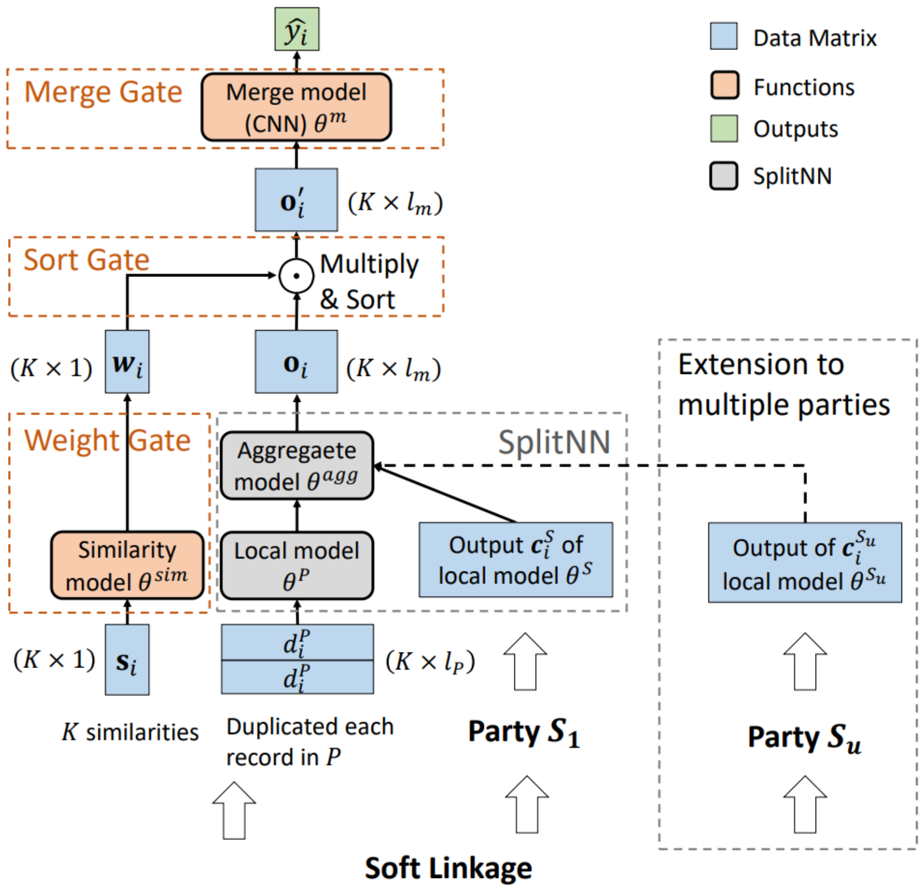
Real estate company & Airbnb
Linked on housing address

Task: Predict housing price

Only linking records with top1 similarity may not capture key features

MOTIVATION

Our Design: FedSim



Weight Gate: grant each record pair a weight according to its similarity

Sort Gate: sort the record pairs according to weights

Merge Gate: a CNN with $n \times 1$ kernel to merge the record pairs with similar weights

Our Design: FedSim

Table 1: Performance on real-world datasets

Algorithms	**house (numeric)** $\Delta = 34.05$	**bike (numeric)** $\Delta = 14.26$	**hdb (numeric)** $\Delta = 20.69$	**game (string)** $\Delta = 4.14$	**company (string)** $\Delta = 10.50$
Solo	58.31±0.28	272.83±1.50	29.75±0.15	85.27±0.29%	42.67±0.66
Exact	-	-	-	89.25±0.12%	44.44±1.95
Top1Sim	58.54±0.35	256.19±1.39	31.56±0.21	92.71±0.08%	42.84±0.77
FeatureSim	66.39±0.15	273.29±0.37	37.39±0.29	91.13±0.23%	39.24±1.80
AvgSim	51.92±0.65	239.85±0.40	34.12±0.19	90.84±0.14%	38.19±0.91
FedSim (w/o Weight)	42.82±0.20	236.79±0.29	27.18±0.08	92.79±0.13%	41.00±1.19
FedSim (w/o Sort)	52.14±0.58	238.30±0.81	36.35±0.42	92.79±0.10%	38.28±1.56
FedSim (w/o CNN)	42.62±0.20	235.97±0.42	27.76±0.13	92.50±0.12%	39.63±1.31
FedSim	**42.12±0.23**	**235.67±0.27**	**27.13±0.06**	**92.88±0.11%**	**37.08±0.61**

FedSim outperforms all the baselines in five real world datasets

Conclusion

FedSim: Coupled Design of Linkage and Training

- Empirical study on real applications in German record linkage center
- Coupled framework of record linkage and VFL training – FedSim
- Metric to estimate the improvement of FedSim w/o training
- Theoretical analysis on the privacy of FedSim
- Experiments on real-world and synthetic datasets

GitHub link: https://github.com/Xtra-Computing/FedSim

THANK YOU