Data Sharing

-- -- From a Federated Learning Perspective

Dandan Li

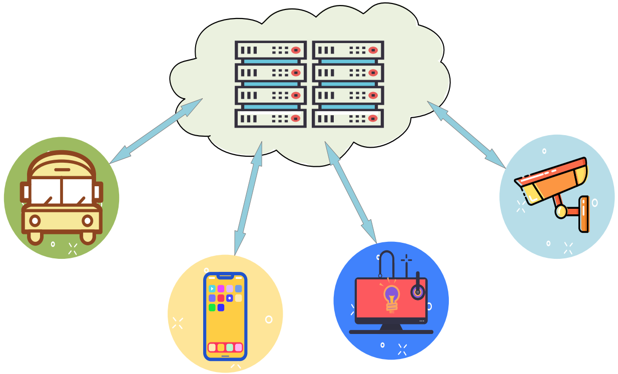
School of Computer (National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications (BUPT)

Outline

outline

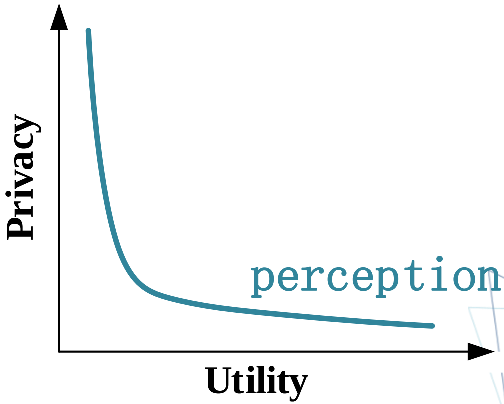
1 Background

- massive users' (private) data + AI spawned many smart industries: smart healthcare, intelligent transport.



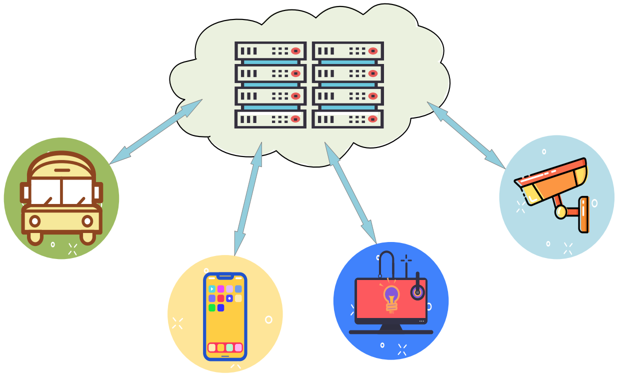
- collect users' (private) data to a central server, which leads to information leakage.

The higher the utility, the worse the privacy.



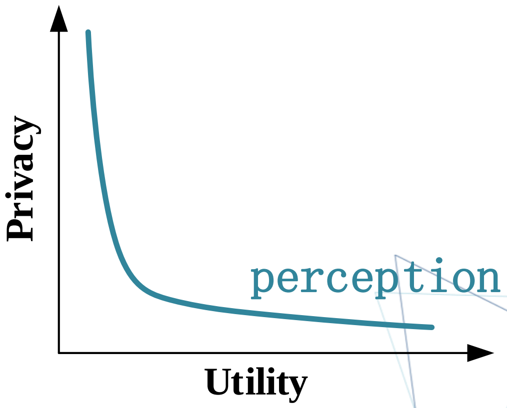
1 Background

- massive user's (private) data + AI spawned many smart industries: smart healthcare, intelligent transport.



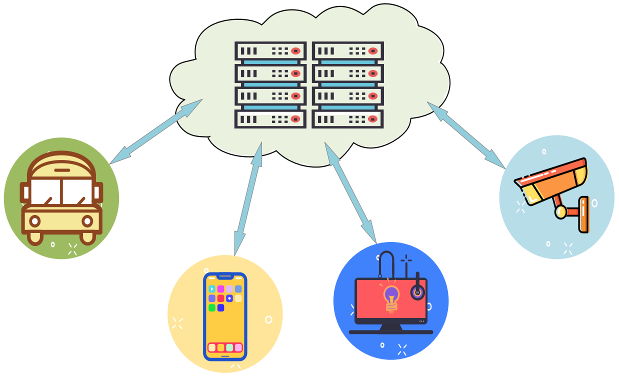
- collect user's (private) data to a central server, which leads to information leakage.

how to balance the utility and privacy ?



1 Background

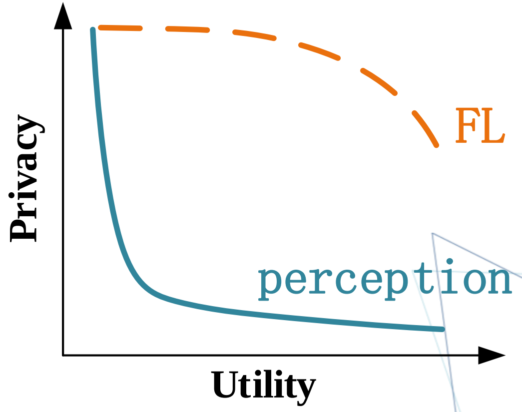
- massive user's (private) data + AI spawned many smart industries: smart healthcare, intelligent transport.



- collect user's (private) data to a central server, which leads to information leakage.

how to balance the utility and privacy ?

federated learning



1 Background

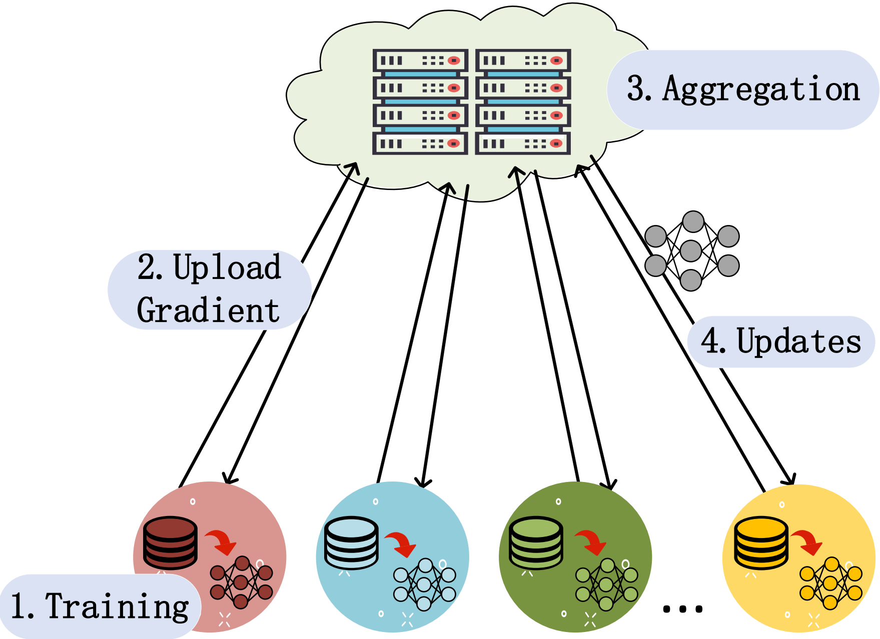
① **Train:** Each client performs model training based on local dataset.

② **Upload:** Each client sends the trained model parameters to server.

③ ***Aggregation:*** Central server aggregates received models.

④ **Update:** The server sends the updated model to each client.

⑤ repeat steps ①-④ until predetermined condition is met.



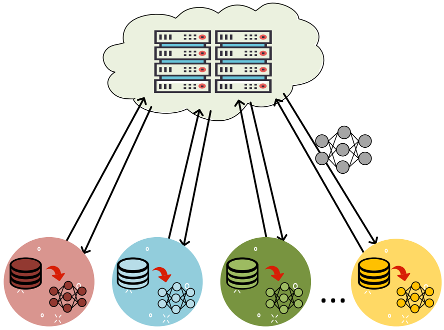
The workflow of federated learning

The raw data doesn't move and the model does.

1 Background

Chanllenges:

- Data and device heterogeneous:
  - Non-IID data
  - Different devices abilities form CPU, memory, disk read and write speed,etc.



- Communication pressure:
  - For server, **models of massive clients are uploaded to the server** (the only aggregated node) , which causes the server to be congested, furter, causes the time of obtained global model to be longer.
  - For clients, **the network states are dynamic and different**, which causes uplink communication time is different, further, causes the time of obtained global model to be longer .

Outline

outline

2.1 A Cluster-Asynchronous Federated Multi-Task Learning

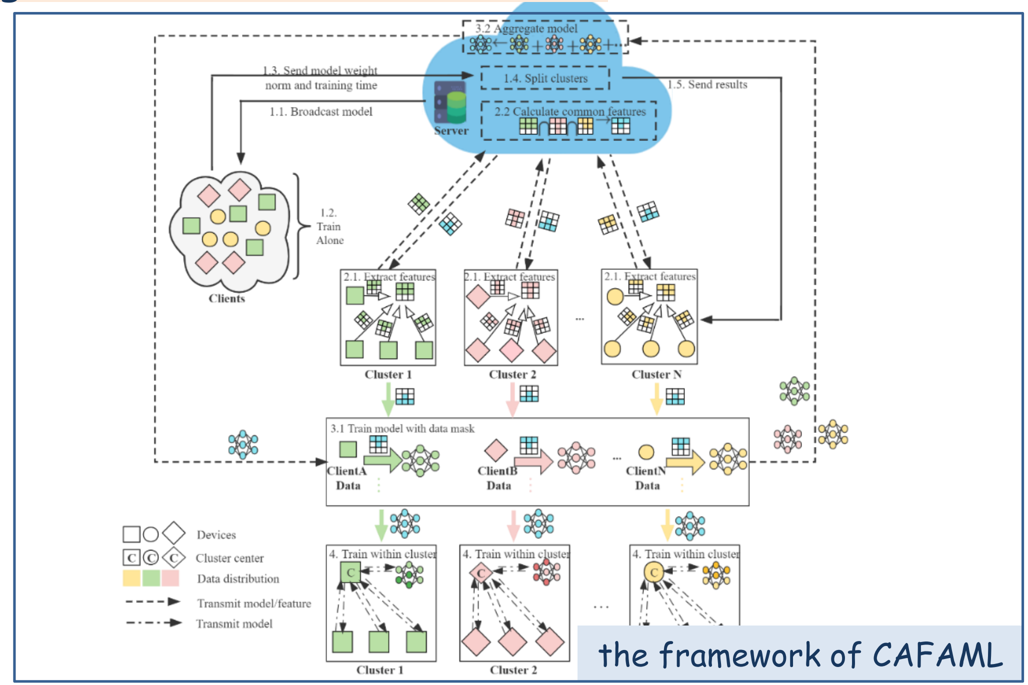
problem 1

Data and device heterogeneous:

- Bad impact on training performance (low model accuracy and long training time).

solution 1

- Cluster based on clients' attributions;
- Extract global-level key features;
- Train global model with feature masking;
- Cluster-Asynchronous.



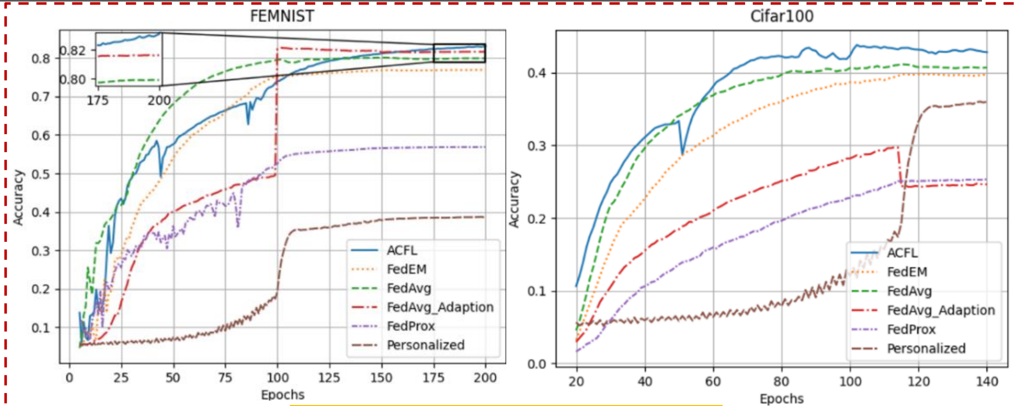
the framework of CAFAML

2.1 A Cluster-Asynchronous Federated Multi-Task Learning

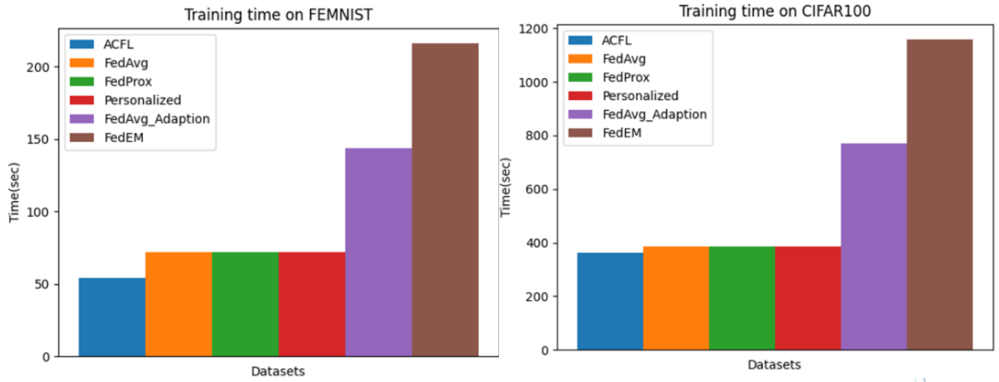
北京郵電大學
Beijing University of Posts and Telecommunications

Performance:

- Datasets
  - FEMNIST
  - CIFAR-100
- Experiment settings
  - Non-iid process:
    - FEMNIST: Natural Non-iid Dataset
    - CIFAR-100: hierarchical Latent Dirichlet Allocation (LDA) process
  - Clients:
    - 539 clients, 120772 samples for FEMNIST
    - 100 clients, 60000 samples for CIFAR-100
  - Devices:
    - Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz
    - Intel(R) Xeon(R) E5-2620 v4 CPU @ 2.10GHz
    - Intel(R) Core (TM) i5-9300H CPU @ 2.40GHz
    - Intel(R) Core (TM) i7-7700HQ CPU @ 2.80GHz
- Metrics
  - Accuracy
  - Training Time.



higher accuracy (blue line)



shorter training time (blue bar)

2.2 Knowledge Distillation with multiple servers in Personalized Federated Learning

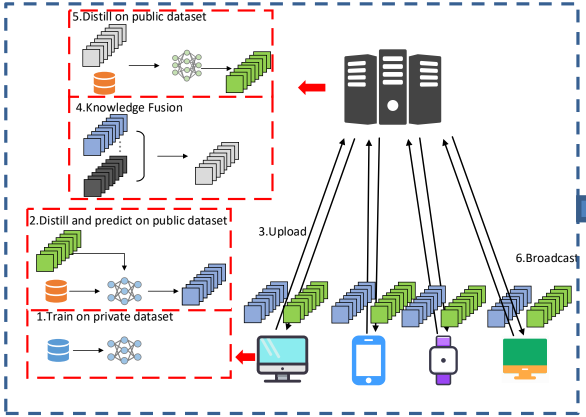
北京郵電大學
Beijing University of Posts and Telecommunications

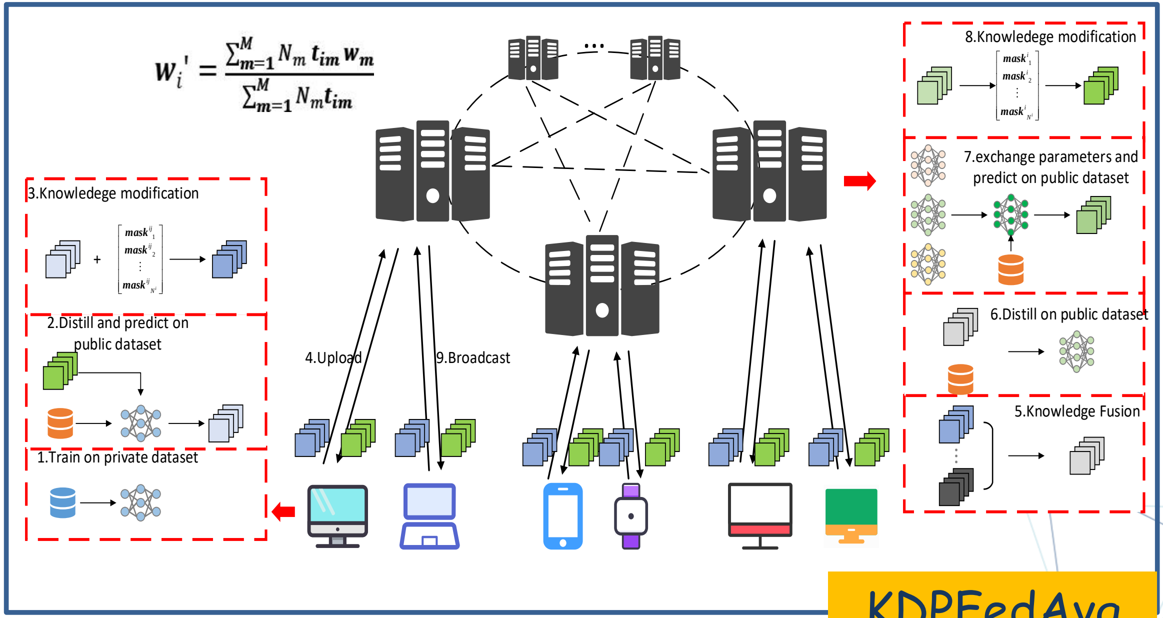
☐ problem 2

- non-IID data
- Communication pressure-- from server

☐ solution 2

- Aggregate model parameters of servers based on topology;
- Federated distillation.





$$w_i^{'} = \frac{\sum_{m=1}^{M} N_m t_{im} w_m}{\sum_{m=1}^{M} N_m t_{im}}$$

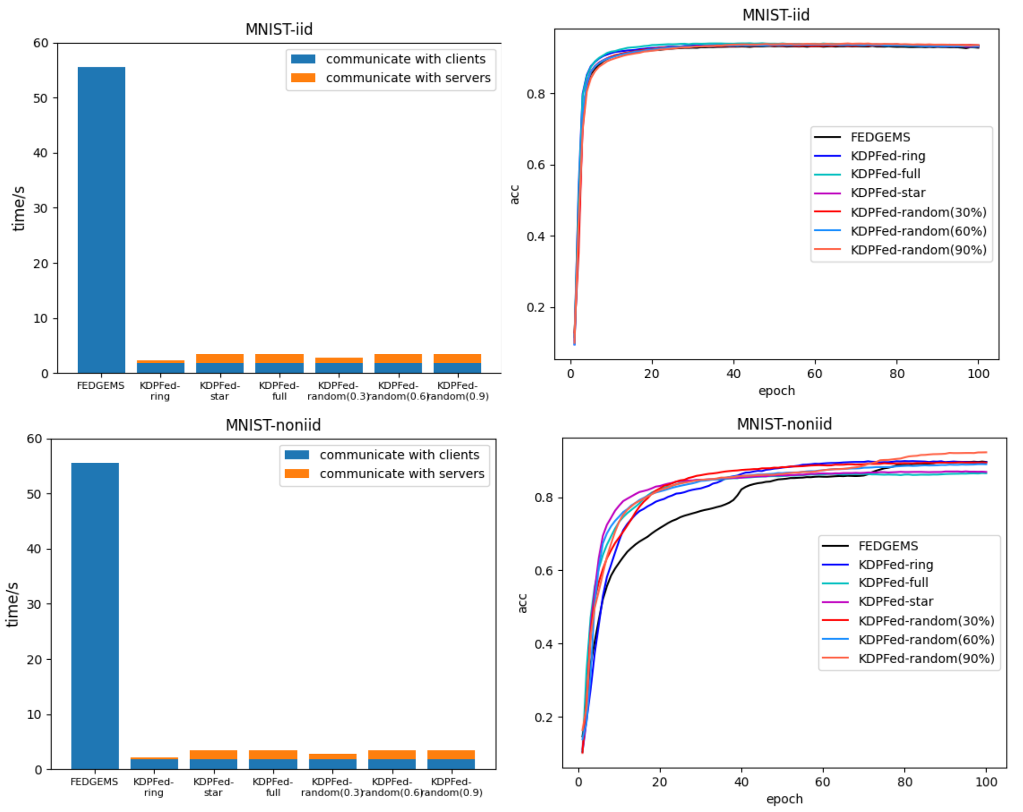
$w_i^{'}$ is the update model parameter of the i_{th} aggregation node
M is the total number of aggregate nodes,
N_m is the number of data of common data set of the aggregate node m,
t_{im} is the value in the topology matrix, which represents the connection relationship between i_{th} node and m_{th} node,
w_m is the model parameter of the m_{th} aggregate node.

KDPFedAvg

2.2 Knowledge Distillation with multiple server in Personalized Federated Learning

Performance

- Datasets
  - MNIST
  - Fashion-MNIST
  - FEMNIST
- Experiment settings
  - Servers: 9
  - Clients: total number is 385
    - Each server randomly generated a certain number of clients: 23, 42, 27, 39, 85, 66, 52, 36, 15
  - Topological type:
    - Ring topology; Fully connected topology; Star topology
    - Random connection topology with probability 30%, 60%, 90%
- Metrics
  - Accuracy
  - Communication time



shorter communication time, similar accuracy

2.3 Communication-Efficient Federated Learning with Adaptive Compression under Dynamic Bandwidth

北京郵電大學 Beijing University of Posts and Telecommunications

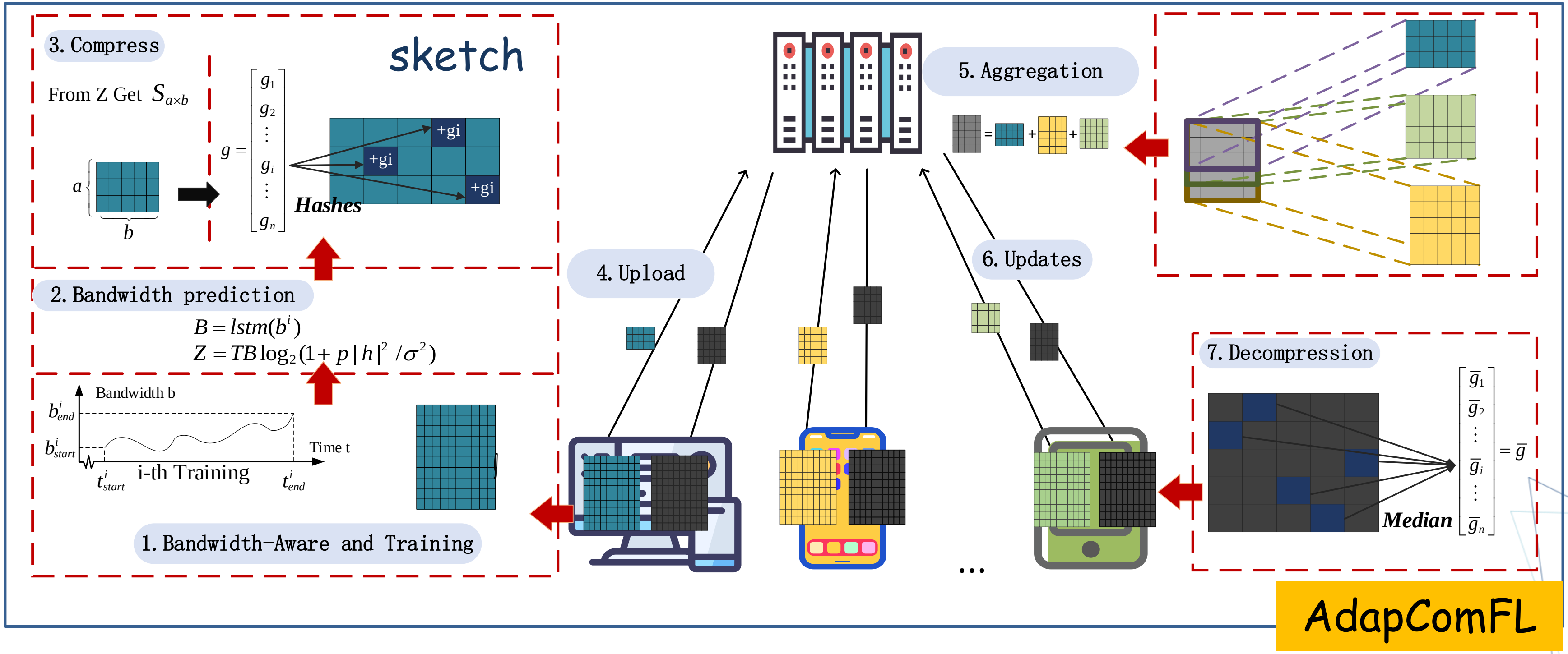
problem 3

Communication pressure-- from clients

- bandwidth is dynamic and different

solution 3

- Aware and predict bandwidth;
- Compress local model adaptively.



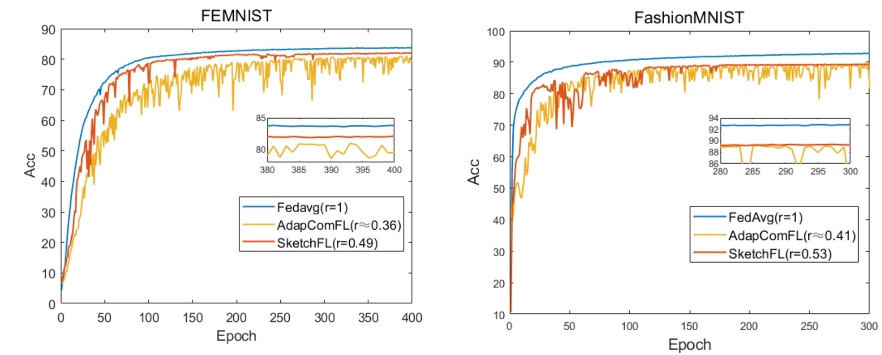
2.3 Communication-Efficient Federated Learning with Adaptive Compression under Dynamic Bandwidth

北京郵電大學
Beijing University of Posts and Telecommunications

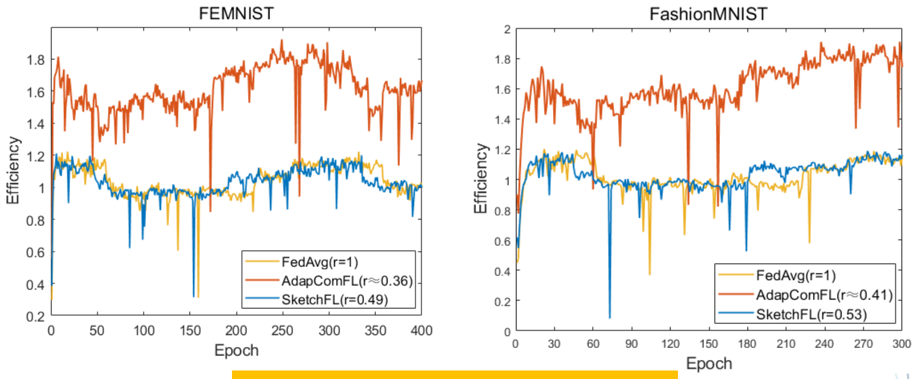
Performance

- **Datasets**
  - Bandwidth datasets:
    we builds a distributed environment to collect bandwidth data
  - Benchmark datasets:
    - FEMNIST
    - Fashion-MNIST
- **Experiment settings**
  - Servers: 1
  - Clients: 7
- **Metrics**
  - Accuracy
  - Communication efficiency
    - Formula: $E = \frac{z}{t}$

    E is communication efficiency
    z is uplink communication data volume
    t is uplink communication delay



competitive accuracy (orange line)



better efficiency (red line)

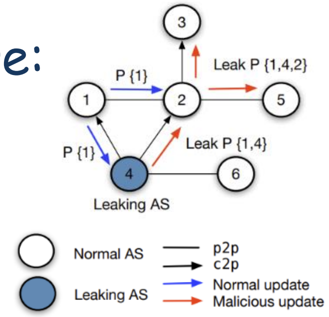
Outline

outline

3.1 Background--BGP Routing Policy

- The Internet is composed of tens of thousands of Autonomous Systems (ASes) and they use Border Gateway Protocol (BGP) to exchange reachability information.

- The routing polices of ASes for path selection are business-oriented.
  - Common business relationship types between ASes are:
    - Customer-to-provider (C2P)
    - Provider-to-customer (P2C)
    - Peer-to-peer (P2P)
  - Common routing policy in the Internet is:
    - routes learned from one peer or provider cannot be propagated to another peer or provider (valley-free rule)



3.1 Background--Route Leaks

- Route leaks occure when an attacker propagates a valid route beyond the scope intended by the routing policy of the involved ASes

  (violate valley-free rule)

  - Causing major outages by redirecting traffic
  - Bring a risk of Man-in-the-Middle attacks

- Main route leak detection methods:
  - Directly sharing routing polices or business relationships (no privacy guarantee)
    - [1-3] add new BGP attribute or extend BGP community to convey business relationship information.
    - IRR[4], registering routing polices on an open database and using the registrations to filter leaks.
    - ASPA[5] adds routing customer-provider objects to RPKI repository.

[1]. Sriram, Kotikalapudi, et al. "Methods for detection and mitigation of bgp route leaks." draft-ietf-idr-route-leak-detection-mitigation-06 (2017).
[2]. Azimov, A., E. Bogomazov, and R. Bush. "Route leak detection and filtering using roles in update and open messages." draft-ymbk-idr-bgp-open-policy-03 (2017).
[3]. Sriram, Kotikalapudi, et al. "Methods for detection and mitigation of bgp route leaks." draft-ietf-idr-route-leak-detection-mitigation-06 (2017).
[4]. Internet Routing Registry (IRR), online. https://www.apnic.net/about-apnic/whois_search/about/what-is-in-whois/irr/
[5]. Azimov, Alexander, et al. "Verification of AS PATH Using the Resource Certificate Public Key Infrastructure and Autonomous System Provider Authorization. IETF, 2018."

3.1 Background--Challenges for detecting route leaks

- ASes are unwilling to reveal their business relationships to others

due to

- Economic issues
- Complexity of routing polices
- ……

3.1 Background--Challenges for detecting route leaks

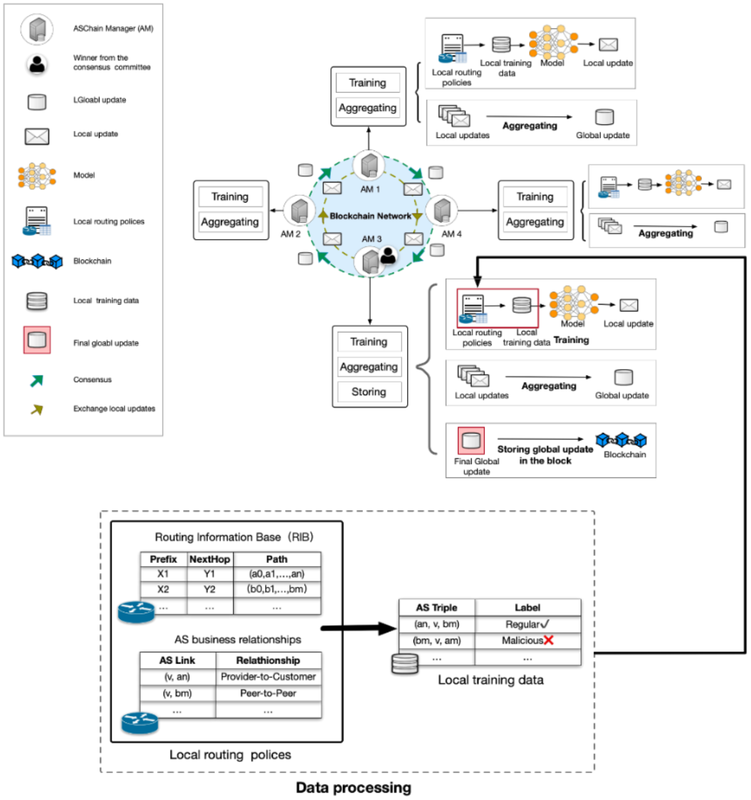
- ASes are unwilling to reveal their business relationships to others

due to

- Economic issues
- Complexity of routing polices
- ……

How to detect route leak while protect business relationship privacy?

3.2 Federated learning route leak detection in Inter-domain routing

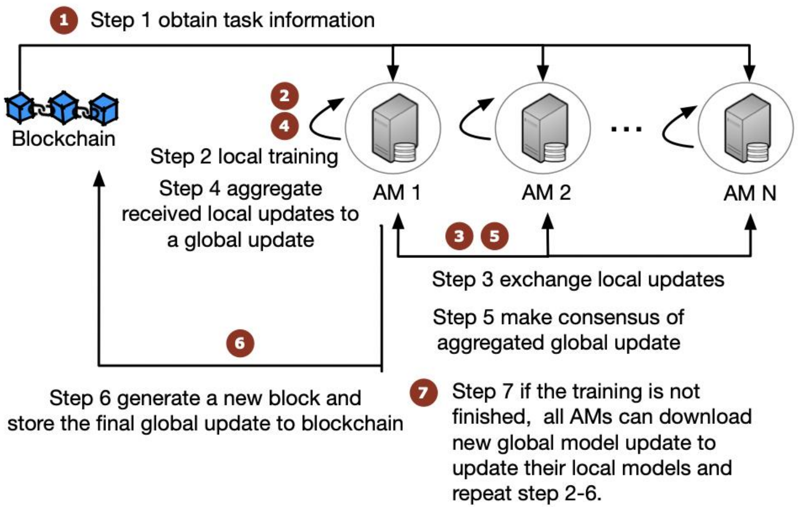


the framework of FL-RLD

- **Aschain Manager**
  - Each AS play roles as client of federated learning and node in blockchain (denoted as AM) .
- **Training Data**
  - Transforming routing policies to AS triples with labels (training datasets)

× instead of directly sharing AS relationships

× labels are generated by valley-free rule using known local routing polices.



the workflow of FL-RLD

- **Step 1** : obtain training task information (i.e., initial model, training epoches) from blockchain.
- **Step 2 to Step 3**: train local model locally and upload local model to blokchain.
- **Step 4 to Step 5:** aggregate all local model and then global update model is obtained
- **Step 6:** the aggerated global update model is stored to blockchain
- **Step 7:** if the training cannot satisfy fixed condition, steps 2-6 are repeated.

3.3 Performance

- Topology

CAIDA IPv6 AS relationship dataset, Jan, 2021

(12,721 ASes, 173,462 AS links)

- Evaluation metrics

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1score = 2\frac{Precision * Recall}{Precision + Recall}$$

3.3 Performance

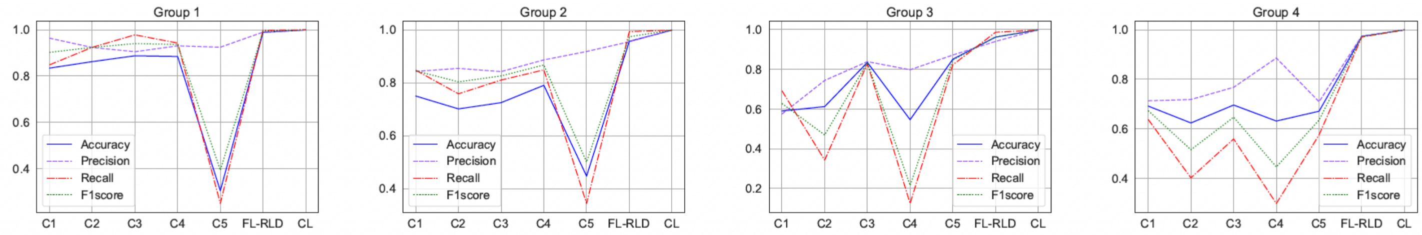
- 4 groups of experiments, each group has 5 clients

TABLE 1: The triple distribution of different groups

	Data size	Anomaly	Regular	Anomaly %	Regular %
Group 1 (unbalanced data size + unbalanced class distribution)	13550	12224	1326	90.21%	9.79%
Client1 (51.19%)	6936	6192	744	89.27%	10.73%
Client2 (30.92%)	4189	3913	276	93.41%	6.59%
Client3 (0.51%)	69	33	36	47.83%	52.17%
Client4 (14.18%)	1922	1680	242	87.41%	12.59%
Client5 (3.20%)	434	406	28	93.55%	6.45%
Group 2 (balanced data size + unbalanced class distribution)	63468	51066	12402	80.46%	19.54%
Client1 (19.77%)	12549	12099	450	96.41%	3.59%
Client2 (20.69%)	13134	12158	976	92.57%	7.43%
Client3 (19.25%)	12218	7606	4612	62.25%	37.75%
Client4 (19.49%)	12369	10205	2164	82.51%	17.50%
Client5 (20.79%)	13198	8998	4200	68.18%	31.82%
Group 3 (unbalanced data size + balanced class distribution)	416348	208174	208174	50.00%	50.00%
Client1 (8.58%)	35712	17856	17856	50.00%	50.00%
Client2 (35.93%)	149580	74790	74790	50.00%	50.00%
Client3 (43.45%)	180904	90452	90452	50.00%	50.00%
Client4 (10.40%)	43316	21658	21658	50.00%	50.00%
Client5 (1.64%)	6836	3418	3418	50.00%	50.00%
Group 4 (balanced data size + balanced class distribution)	17090	8512	8578	49.81%	50.19%
Client1 (20%)	3418	1761	1657	51.52%	48.48%
Client2 (20%)	3418	1672	1746	48.92%	51.08%
Client3 (20%)	3418	1724	1694	50.44%	49.56%
Client4 (20%)	3418	1679	1739	49.12%	50.88%
Client5 (20%)	3418	1676	1742	49.04%	50.97%

3.3 Performance

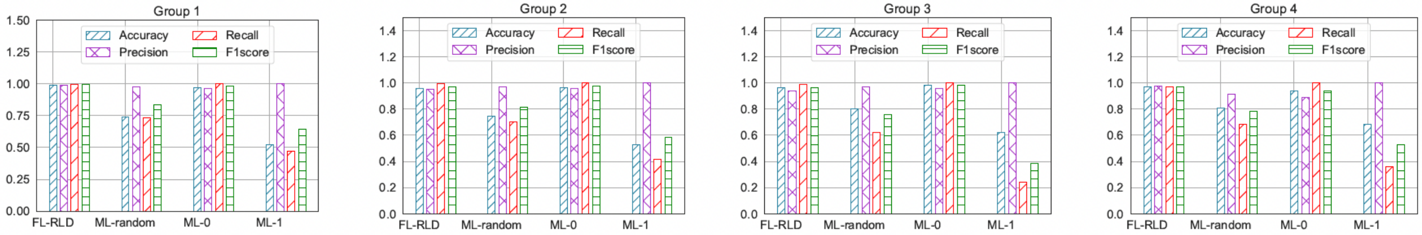
single AS vs. FL-RLD



(a) Unbalanced data size + unbalanced class distribution (b) Data size balance + unbalanced class distribution (c) Unbalanced data size + balanced class distribution (d) Data size balance + balanced class distribution

Fig. 4: The Performance of FL-RLD method compared with single AS learning method (C1, C2, C3, C4, C5) and Central Learning (CL) method

global repository vs. FL-RLD

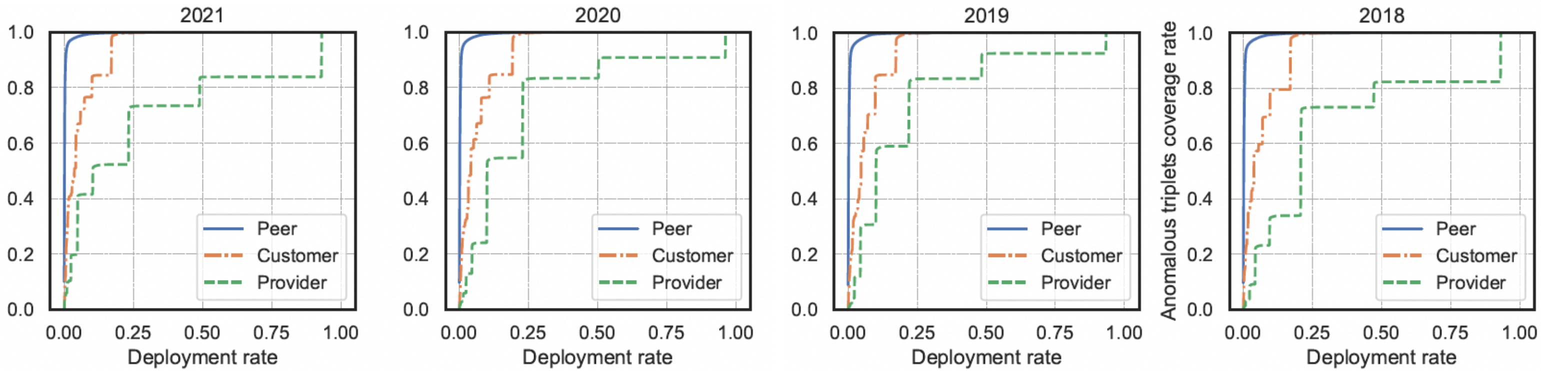


(a) Anomaly 90.21% vs Regular 9.79% (b) Anomaly 80.46% vs Regular 19.54% (c) Anomaly 50% vs Regular 50% (d) Anomaly 49.81% vs Regular 50.19%

Fig. 5: The performance comparison of FL-RLD and other methods.

北京郵電大學 Beijing University of Posts and Telecommunications

Deployment strategies



The more number of malicious triples,the better detection result.

Peer deployment strategy can cover the most number of malicious triples than other two strategies with the same deployment rate.

ASes with a large number of peers can be deployed which achieves better detection results.

Outline

北京邮電大學
Beijing University of Posts and Telecommunications

outline

1. Background
2. Related research
3. Federated route leak detection
4. Conclusion

4. Conclusion

- faced with **data heterogeneous + device heterogeneous**, CAFAML achieves higher accuracy and shorter training time.
- faced with **data heterogeneous +communication pressure**, KDPFedAvg achieves shorter communicaition time with similar accuracy.
- faced with **communication pressure from clients**, AdapComFL achieves better communicaition efficientcy with competitive accuracy.
- for route leak detection， deployment Suggestion of FL-RLD： ASes with a large number of peers can be deployed which achieve better detection results.

Thank you

dandan Li
dandl@bupt.edu.cn