# Federated Graph Classification over Non-IID Graphs

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Federated learning has emerged as an important paradigm of training machine learning models in different domains. For graph-level tasks such as graph classification, graphs can also be regarded as a special type of data samples, which can be collected and stored in separate local systems. Similar to other domains, multiple local systems, each holding a small set of graphs, may benefit from collaboratively training a powerful graph mining model, such as the popular graph neural networks (GNNs). To provide more motivation towards such endeavors, we analyze real-world graphs from different domains to confirm that they indeed share certain graph properties that are statistically significant compared with random graphs. However, we also find that different sets of graphs, even from the same domain or same dataset, are non-IID regarding both graph structures and node features. To handle this, we propose a graph clustering federated learning (GCFL) framework that dynamically finds clusters of local systems based on the gradients of GNNs, and theoretically justify that such clusters can reduce the structure and feature heterogeneity among graphs owned by the local systems. Moreover, we observe the gradients of GNNs to be rather fluctuating in GCFL which impedes high-quality clustering, and design a gradient sequence-based clustering mechanism based on dynamic time warping (GCFL+). Extensive experimental results and in-depth analysis demonstrate the effectiveness of our proposed frameworks.

## 1 Introduction

Federated learning (FL) as a distributed learning paradigm that trains centralized models on decentralized data has attracted much attentions recently [26, 44, 23, 17, 16]. FL allows local systems to benefit from each other while keeping their own data private. Especially, for local systems with scarce training data or lack of diverse distributions, FL provides them with the potentiality to leverage the power of data from others, in order to facilitate the performance on their own local tasks. FL concerns many problems including data privacy, data heterogeneity, communication efficiency, etc. One major problem from the data perspective is data distribution heterogeneity, since the decentralized data, collected by different institutes using different methods and aiming for different tasks, are highly likely to follow non-identical distributions. Prior works approach this problem from different aspects, including optimization process [23, 17], personalized FL [12, 5, 7], clustered FL [9, 14, 2], etc.

As more advanced techniques are developed for learning with graph data, using graphs to model and solve real-world problems becomes more popular. One important scenario of graph learning is graph classification, where models such as graph kernels [41, 32, 31, 33, 1] and graph neural networks [20, 40, 42] are used to predict graph-level labels based on the node features and link structures in the graph-level samples. One real scenario of graph classification is molecular property or activity prediction, which is an important task in cheminformatics and AI medicine like de novo drug design. In the area of bioinformatics, graph classification can be used to learn the representation of proteins and classify them into enzymes or non-enzymes. For social networks, e.g., a collaboration network, graph classification can learn from its sub-networks about the information of studying areas, topics, genre, etc. More applicable scenarios include geographic networks, temporal networks, etc.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Since the key idea of FL is the sharing of underlying common information, as [22] discusses that real-world graphs preserve many common properties, we become curious about the question, *whether real-world graphs from heterogeneous sources (e.g., different datasets or even divergent domains) can provide useful common information among each other?* To understand this question, we first conduct preliminary data analysis to explore real-world graph properties, and try to find clues about common patterns shared among graphs across datasets. As shown in Table 1, we analyze four typical datasets from different domains, i.e., NCI1 (molecular structures), ENZYMES (protein structures) and IMDB-BINARY (social communities), and MSRC_21 (superpixel networks). We find them to indeed share certain properties that are statistically significant compared to random graphs with the same sizes (generated with the Erdős–Rényi model [6, 10]). Such observations confirm the claim about common patterns underlying real-world graphs, which can largely influence the graph mining models and motivates us to consider the FL of graph classification across datasets and even domains.

Table 1: Data analysis on important graph properties shared among real-world graphs across different domains. For example, large Kurtosis values [29] indicate long-tail distribution of node degrees, which is observed in ENZYMES, IMDB-BINARY and MSRC_21; similar average shortest path lengths are observed in NCI1, ENZYMES and MSRC_21, although their actual graph sizes are rather different; large LC and CC are observed in almost all graphs.

| Property | kurtosis of degree distribution | | | avg. shortest path length | | | largest component size (LC, %) | | | clustering coefficient (CC) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | real | random | p-value | real | random | p-value | real | random | p-value | real | random | p-value |
| PTC_MR (chemical) | 2.1535 | 2.0762 | 0.1215 | 3.36 | 1.48 | $\sim 0$ | 100 | 32.32 | $\sim 0$ | 0.0095 | 0.0068 | 0.4718 |
| ENZYMES (biological) | 3.0075 | 2.7543 | 0.0001 | 4.44 | 1.77 | $\sim 0$ | 98.25 | 99.50 | $\sim 0$ | 0.4564 | 0.2941 | $\sim 0$ |
| IMDB-BINARY (social) | 8.9262 | 2.7910 | $\sim 0$ | 1.48 | 2.10 | $\sim 0$ | 100 | 36.50 | $\sim 0$ | 0.9471 | 0.0120 | $\sim 0$ |
| MSRC_21 (visual) | 3.6959 | 2.9329 | $\sim 0$ | 4.09 | 2.35 | $\sim 0$ | 100 | 99.94 | $\sim 0$ | 0.5147 | 0.0986 | $\sim 0$ |

Although common patterns exist among graph datasets, we can still observe certain heterogeneity. In fact, the detailed graph structure distributions and node feature distributions can both diverge due to various reasons. To demonstrate this, we design and evaluate a structure heterogeneity measure and a feature heterogeneity measure in different scenarios (c.f. Section 4.1). We refer to the graphs possibly with significant heterogeneity in our cross-dataset FL setting as non-IID graphs, which concerns both structure non-IID and feature non-IID, where naïve FL algorithms like `FedAvg` [26] can fail and even backfire (c.f. Section 6.2). Moreover, as the heterogeneity varies from case to case, a dynamic FL algorithm is needed to keep track of such heterogeneity of non-IID graphs while conducting collaborative model training.

Due to the observations that the graphs in one client can be similar to those in some clients but not the others, we get motivated by [2] and find it intuitive to consider a clustered FL framework, which assigns local clients to multiple clusters with less data heterogeneity. To this end, we propose a novel graph-level clustered FL framework (termed GCFL) through integrating the powerful GIN model [40] into clustered FL, where the server can dynamically cluster the clients based on the gradients of GIN without additional prior knowledge, while collaboratively training multiple GINs as necessary for homogeneous clusters of clients. We theoretically analyze that the model parameters of GIN indeed reflect the structures and features of graphs, and thus using the gradients of GIN for clustering indeed yields clusters with reduced heterogeneity of both structures and features.

Although GCFL can theoretically achieve homogeneous clusters, during its training, we observe that the gradients transmitted at each communication round fluctuate a lot (c.f. Section 5.1), which could be caused by the complicated interactions among clients regarding both structure and feature heterogeneity, making the local gradients towards divergent directions. In the vanilla GCFL framework, the server calculates a matrix for clustering only based on the last transmitted gradients, which ignores the client's multi-round behaviors. Therefore, we further propose an improved version of GCFL with gradient-series-based clustering (termed GCFL+).

We conducted extensive experiments with various settings to demonstrate the effectiveness of our frameworks. Moreover, we provide in-depth analysis on the capability of them on reducing both structure and feature heterogeneity of clients through clustering. Lastly, we analyze the convergence of our frameworks. The experimental results show surprisingly positive results brought by our novel setting of cross-dataset/cross-domain FL for graph classification, where our GCFL+ framework can effectively and consistently outperform all baselines.

## 2 Related works

**Federated Learning** Federated learning has gained increasing attention as a training paradigm under the setting where data distributed at remote devices and collaboratively trained under the coordination of a central server. `FedAvg` was first proposed by [25] which illustrates the general setting of Federated Learning framework. Since federated learning relies on the optimization by SGD, data non-IID distribution will not guarantee the stochastic gradients to be an unbiased estimation of the full gradient. Thus it will hurt the convergence of the federated learning algorithms. Multiple experiments ([44, 23, 17]) have shown that convergence will be slow and unstable, and the accuracy will degrade with `FedAvg` when data at each client are statistically heterogeneous (non-IID). ([44, 14, 11]) proposed different data sharing strategies to tackle the data heterogeneity problem by sharing the local device data or server-side proxy data. Several works explored the convergence guarantee under the data non-IID setting by assuming bounded gradients ([36, 43]), or viewing as additional noise ([18]). There are also works seeking to reduce the variance of the clients ([24, 17, 23]). Furthermore, multiple works has been proposed which explore the connection between model-agnostic meta-learning (MAML) and personalized federated learning [8, 3].

**Federated Learning on Graphs** Although federated learning has been thoroughly studied with the Euclidean datasets, there are few studies about graph based federated learning. [21] proposed the first work that applied the federated learning paradigm to the graph learning. [4] proposed a generalized federated knowledge graph embedding framework that can be applied for multiple knowledge graph embedding algorithms. Moreover, there are several works exploring the Graph Neural Networks (GNNs) under the federated learning setting. [15, 45, 37] focused on tackling the privacy issue of federated GNN. [34] incorporates model-agnostic meta-learning (MAML) into federated learning framework, which enables non-IID data while also preserves the model generalizability. [35] proposes a computationally efficient way of GCN architecture search with federated learning.

## 3 Preliminaries

### 3.1 Graph Neural Networks (GNNs)

[39] provides a taxonomy that categorizes Graph Neural Networks (GNNs) into recurrent GNNs, convolutional GNNs, graph autoencoders, and spatial-temporal GNNs. Generally, given the structure and feature information of a graph $G = (V, E, X)$, where $V$, $E$, $X$ denotes nodes, links and node features, GNNs target to learn the representations of graphs, such as a node embedding $h_v \in \mathbb{R}^{d_v}$, or a graph embedding $h_G \in \mathbb{R}^{d_G}$. A GNN typically consists of message propagation and neighborhood aggregation, in which each node iteratively gathers the information propagated by its neighbors, and aggregates them with its own information to update its representation. Generally, an $L$-layer GNN can be formulated as

$$h_v^{(l+1)} = \sigma(h_v^{(l)}, agg(\{h_u^{(l)}; u \in \mathcal{N}_v\})), \forall l \in [L], \tag{1}$$

where $h_v^{(l)}$ is the representation of node $v$ at the $l^{th}$ layer, and $h_v^{(0)} = x_v$ is the node feature. $\mathcal{N}_v$ is neighbors of node $v$, $agg(\cdot)$ is a aggregation function that can vary for different GNN variants, and $\sigma$ represents a activation function.

For a graph-level representation $h_G$, it can be pooling from the representations of all nodes, as

$$h_G = pool(\{h_v; v \in V\}). \tag{2}$$

For GNNs on graph-level tasks, a graph aggragation layer with different $readout(\cdot)$ functions on the top of a GNN is often applied, in order to aggregate the embeddings of all nodes on the graph into a single emebdding vector to achieve tasks like graph classification and regression.

### 3.2 The `FedAvg` algorithm

McMahan et al. [26] proposed a SGD-based aggregating algorithm, `FedAvg`, based on the fact that SGD is widely used and powerful for optimization. `FedAvg` is the first basic federated learning algorithm and is commonly used as the start point for more advance FL framework design.

The key idea of `FedAvg` is to aggregate the updated model parameters transmitted from local clients and then re-distributed the averaged parameters back to each client. In details, suppose $m$ clients in total, at each communication round $t$, the server firstly samples a partition of clients $\{\mathbb{S}_i\}^{(t)}$, and for each client $\mathbb{S}_i$ in the $\{\mathbb{S}_i\}^{(t)}$, it trains the model downloaded from the server locally with its own data distribution $\mathcal{D}_i$ for $E$ epochs. The client $\mathbb{S}_i$ then transmits its updated parameters $w_i^{(t)}$ to the server, and the server will aggregate these updates by

$$w^{(t+1)} = \sum_{i=1}^{m} \frac{|D_i|}{|D|} w_i^{(t)}, \tag{3}$$

where $|D_i|$ is the size of data samples in client $\mathbb{S}_i$ and $|D|$ is the total size of samples over all clients. After generating the aggregated parameters (the global model updates), the server broadcasts the new parameters $w^{(t+1)}$ to remote clients, and at the $(t+1)$ round clients use $w^{(t+1)}$ to start for another $E$ epochs of their local training.

# 4 The `GCFL` framework

## 4.1 Non-IID structures and features across clients

From Table 1 we notice that real-world graphs tend to share certain general properties across different graphs, datasets and even domains, which motivates the graph-level federated learning framework. However, there still exist differences when the detailed graph structures and node features are being considered. In Table 2, we present the average pair-wise structure heterogeneity and feature heterogeneity among graphs in a single dataset, a single domain, and across different domains. Specifically, for structure heterogeneity, we use the Anonymous Walk Embeddings (AWEs) [13] to generate a representation for each graph, and compute the Jensen-Shannon distance between the AWEs of each pair of graphs; for feature heterogeneity, we calculate the empirical distribution of feature similarity between all pairs of linked nodes in each graph, and compute the Jensen-Shannon divergence between the feature similarity distributions of each pair of graphs.

As we can observe in Table 2, both graph structures and features demonstrate different levels of heterogeneity within a single dataset, a single domain, and across different domains. We refer to graphs with such structure and feature heterogeneity as non-IID graphs. Intuitively, directly applying naïve federated learning algorithms like `FedAvg` on clients with non-IID graphs can be ineffective and even backfiring. To be specific, structure heterogeneity makes it difficult for a model to capture the universally important graph structure patterns across different clients, whereas feature heterogeneity makes it hard for a model to learn the universally appropriate message propagation functions across different clients. How can we leverage the shared graph properties among clients while addressing the non-IID structures and features across clients?

## 4.2 Problem formulation

Motivated by our real graph data analysis in Tables 1 and 2, we propose a novel framework of **G**raph **C**lustered **F**ederated **L**earning (GCFL). The main idea of GCFL is to jointly find clusters of clients with graphs of similar structures and features, and train the graph mining model with `FedAvg` among clients in the same clusters.

Specifically, we are inspired by the Clustered Federated Learning (CFL) framework on Euclidean data [30] and consider a clustered FL setting where there is one central server and a set of $n$ local clients $\{\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_n\}$. Different from the traditional FL setting, the server can dynamically cluster the clients into a set of clusters $\{\mathbb{C}_1, \mathbb{C}_2, \ldots\}$ and maintain $m$ cluster-wise models. In our GCFL setting, each local client $\mathbb{S}_i$ owns a set of graphs $\mathcal{G}_i = \{G_1, G_2, \ldots\}$, where each $G_j = (V_j, E_j, X_j, y_j)$ is a graph data sample with a set of nodes $V_j$, a set of edges $E_j$, node features $X_j$, and a graph class label $y_j$. The task on each local client $\mathbb{S}_i$ is graph classification that predicts the class label $\hat{y}_j = h_k^*(G_j)$ for each graph $G_j \in \mathcal{G}_i$, where $h_k^*$ is the collaboratively learned optimal graph mining model for cluster $\mathbb{C}_k$ to which $\mathbb{S}_i$ belongs. Our goal is to minimize the loss function $F(\Theta_k) := \mathrm{E}_{\mathbb{S}_i \in \mathbb{C}_k}[f(\theta_{k,i}; \mathcal{G}_i)]$, for all clusters $\{\mathbb{C}_k\}$. The function $f(\theta_{k,i}; \mathcal{G}_i)$ is a local loss function for client $\mathbb{S}_i$ which belongs to cluster $\mathbb{C}_k$. In the meantime, we also aim to maintain a dynamic cluster assignment $\Gamma(\mathbb{S}_i) \rightarrow \{\mathbb{C}_k\}$ based on the federated learning process.

Table 2: Summary of the average heterogeneity of features and structures for some datasets. In general, the structure heterogeneity increases from the settings of one dataset to across-dataset, and to across-domain. However, the feature heterogeneity is more case-by-case, and the high variances indicate that graphs could have large feature divergence even within the same dataset. Additionally, it is not necessarily true that one dataset itself should be more homogeneous (e.g., IMDB-BINARY).

| dataset 1 | IMDB-BINARY (social) | COX2 (molecules) | COX2 (molecules) | COX2 (molecules) | COX2 (molecules) |
|---|---|---|---|---|---|
| dataset 2 | IMDB-BINARY (social) | COX2 (molecules) | PTC_MR (molecules) | ENZYMES (bioinfo) | IMDB-BINARY (social) |
| avg. struc. hetero. | 0.4406 ($\pm$0.0397) | 0.3246 ($\pm$0.0145) | 0.3689 ($\pm$0.0540) | 0.5082 ($\pm$0.0399) | 0.6079 ($\pm$0.0331) |
| avg. feat. hetero. | 0.1785 ($\pm$0.1226) | 0.0427 ($\pm$0.0314) | 0.1837 ($\pm$0.1065) | 0.1912 ($\pm$0.1000) | 0.1642 ($\pm$0.1006) |

### 4.3 Technical design

GNNs are demonstrated to be powerful for learning graph representations and have been wildly used in graph mining. More importantly, the model parameters and their gradients of GNNs can reflect the graph structure and feature information (more details in Section 4.4. Thus, we use GNNs as the graph mining model in our GCFL framework.

Specifically, our GCFL framework leverages the transmitted gradients $\{\Delta\theta_i\}_{i=1}^n$ of clients to dynamically cluster them, in order to maximize the collaboration among more homogeneous clients and eliminate the harm from heterogeneous clients. According to [30], if the data distribution of clients are highly heterogeneous, FL cannot jointly optimize all local loss functions, which means that the norm of gradients are greater than zero. Here, we introduce a hyper-parameter $\epsilon_1$ as a criterion to check for stopping the general FL stage in GCFL algorithm, if

$$\delta_{mean} = \| \sum_{i\in[n]} \Delta\theta_i \| < \epsilon_1. \tag{4}$$

In the meantime, if some gradients have a large norm, it means that they fail to approach to their stationary points, and thus clustering is needed to eliminate the negative influence among heterogeneous clients. We then introduce the second criterion with a hyper-parameter $\epsilon_2$ to split the clusters when

$$\delta_{max} = \max(\|\Delta\theta_i\|) > \epsilon_2 > 0. \tag{5}$$

The GCFL framework follows a top-down bi-partitioning mechanism. At each communication round $t$, the server receives $m$ sets of gradients $\{\{\Delta\theta_{i_1}\}, \{\Delta\theta_{i_2}\}, \ldots, \{\Delta\theta_{i_m}\}\}$ from clients in clusters $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_m\}$. For a cluster $\mathbb{C}_k$, if $\delta^k_{mean}$ and $\delta^k_{max}$ satisfy the Eqs. 4 and 5, the server will calculate a cluster-wise cosine similarity matrix $\alpha_k$, and use it to perform an agglomerative clustering, which divides the cluster $\mathbb{C}_k \rightarrow \{\mathbb{C}_{k1}, \mathbb{C}_{k2}\}$. The clustering mechanism based on Eqs. 4 and 5 can automatically and dynamically determine the number of clusters along the FL, while the two hyper-parameters of $\epsilon_1$ and $\epsilon_2$ can be easily set through some simple experiments on the validation sets following [30].

For a client $\mathbb{S}_i$ in cluster $\mathbb{C}_k$, it tries to find $\hat{\theta}_{k,i}$ that is close to the real solution $\theta^*_{k,i} = \arg\min_{\theta_i\in\Theta_k} f(\theta_{k,i}; \mathcal{G}_i)$. At a communication round $t$, the client $\mathbb{S}_k$ transmits its gradient to the server

$$\Delta\theta^t_{k,i} = \hat{\theta}^t_{k,i} - \theta^{t-1}_{k,i}. \tag{6}$$

Since the server maintains the cluster assignments, it can aggregate the gradients cluster-wise by

$$\theta^{t+1}_k = \theta^t_k + \sum_{i\in[n_k]} \Delta\theta^t_{k,i}. \tag{7}$$

### 4.4 Theoretical analysis

We investigate the problem of federated graph learning with multi-domain data distribution, and use the gradient-based federated learning paradigm [25] to facilitate the model training. We theoretically prove that the gradient-based FL algorithm on GNNs can capture the structure and feature heterogeneity, along with the task difference between data from different domains. We study two general problems in order to prove that the gradient can reflect the feature, structure, and task information.

**Definition 4.1** *Let a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ which maps from the metric space $(\mathcal{X}, d)$ to $(\mathcal{Y}, d')$, the function $f$ is considered to have $\delta$ distortion if $\forall u, v \in \mathcal{X}$, $\frac{1}{\delta} d(u, v) \leq d'(f(u), f(v)) \leq d(u, v)$.*

**Theorem 4.1** *(Bourgain theorem) Given an $n$-point metric space $(\mathcal{X}, d)$ and an embedding function $f$ as defined above, then $\forall u, v \in \mathcal{X}$, there exist an embedding mapped from $(\mathcal{X}, d)$ to $\mathbb{R}^k$ with the distortion of the embedding being $O(\log n)$.*

**Problem 1.** *GCFL which involves the communication of the gradients between graphs with heterogeneous structures distributed among different clients, the structure and feature difference can be captured by the GNN gradients.*

For simplicity, we solve Problem 1 with the GNN of Simple Graph Convolutions (SGC) [38], through the following two propositions.

**Proposition 4.1** *Given a graph $G$ with fixed structure represented by the normalized graph Laplacian $\mathcal{L} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, feature represented with $X$, and an SGC $f(\mathcal{L}, X) = softmax(\mathcal{L}^K X\Theta)$ with weights $\Theta$ trained on graph $G$. If we have another graph $G'$ with different structure $\mathcal{L}'$, the weight difference $||\Theta' - \Theta||_2$ is bounded with the structure difference.*

**Proposition 4.2** *Given a graph $G$ with fixed structure represented by the normalized graph Laplacian $\mathcal{L} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, feature represented with $X$, and an SGC $f(\mathcal{L}, X) = softmax(\mathcal{L}^K X\Theta)$ with weights $\Theta$ trained on graph $G$. If we have another graph $G'$ with different feature $\mathcal{X}'$, the weight difference $||\Theta' - \Theta||_2$ is bounded with the feature difference.*

We prove proposition 4.1 and 4.2 in the Appendix. We use the *Bourgain theorem* to bound the difference between embeddings generated with different graph structures/features, and prove that the feature and structure information of a graph is incorporated into the model weights (gradients). By proving that the model weights (gradients) are bounded with the structure/feature difference, we show that the gradient will change with the structure and feature. This further justifies that our proposed gradient based clustering framework GCFL is able to capture the structure and feature information. In addition, we also study the following problem, which allows our GCFL framework to be further extended to cross-task graph-level federated learning in the future.

**Problem 2.** *The communicated gradients in GCFL can also capture the task heterogeneity.*

**Proposition 4.3** *Given a graph $G$ with structure represented by the normalized graph Laplacian $\mathcal{L} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, and feature represented with $X$, if trained with different tasks, we will get the Simple SGC with bounded weights.*

The proof of proposition 4.3 can be found in Appendix.

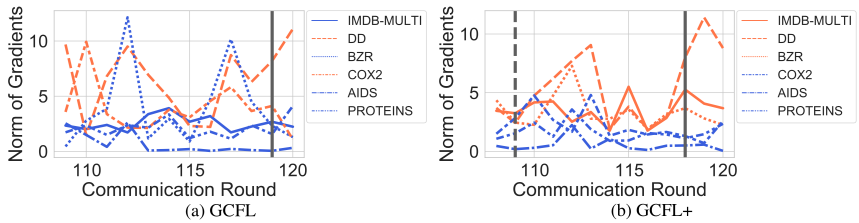
# 5 GCFL+: improved GCFL based on observation sequences of gradients

## 5.1 Fluctuation of gradient norms

When observing the norm of gradients for each communication round in GCFL, as shown in Figure 1, we notice that: 1) the norm of gradients continuously fluctuates; 2) different clients can have divergent scales of gradient norms. The fluctuation of gradient norms and different scales indicate that the updating directions and distances of gradients for clients are diverse, which manifests the structure and feature heterogeneity in our setting again. In our vanilla GCFL framework, the server calculates a cosine similarity matrix based on the last transmitted gradient updates once the clustering criteria are satisfied. However, with the observation that the norm of gradients fluctuates along the communication round, albeit with the constraints of clustering criteria, GCFL clustering based on gradient-point could omit important client behaviors and be misled by noises. For example, in Figure 1 (a), GCFL performs clustering at round 119 based on the gradients at that round, which does not effectively find graphs with lower heterogeneity.

## 5.2 Technical design

Motivated by these observations, we proposed an improved version of GCFL, named GCFL+, which conducts clustering by taking series of gradient norms into consideration. In the GCFL+ framework, the server maintains a multi-variant time-series matrix $Q \in \mathbb{R}^{\{n,d\}}$, where $n$ is the number of clients




(a) GCFL (b) GCFL+

Figure 1: Norm of gradients versus communication round with six clients across datasets.

and $d$ is the length of a gradient series being tracked. At each communication round $t$, the server updates $Q$ by adding in the norm of gradients $\|\Delta\theta_i^t\|$ to $Q(i,:) \in \mathbb{R}^d$ and remove the out-of-date one. GCFL+ uses the same clustering criteria as GCFL (Eqs. 4 and 5). If the clustering criteria are satisfied, the server will calculate a distance matrix $\beta$ in which each cell is the pair-wise distance of two series of gradients. Here, we use a technique called dynamic time warping (DTW) [28] to measure the similarity between two data sequences. For a cluster $\mathbb{C}_k$, the server calculates its distance matrix as

$$\beta_k(p,q) = dist(Q(p,:), Q(q,:)), p, q \in idx(\{\mathbb{S}_i\}), \tag{8}$$

where $idx(\{\mathbb{S}_i\})$ is the indices of all clients $\{\mathbb{S}_i\}$ in cluster $\mathbb{C}_k$. With the distance matrix $\beta$, the server can perform bi-partitioning for clusters who meet the clustering criteria. As a result, in Figure 1 (b), GCFL+ performs clustering at round 118 based on the gradient sequence of length 10, which captures the longer-range behaviors of clients and effectively more homogeneous clusters.

# 6 Experiments

## 6.1 Experimental settings

**Datasets** We use a total of 13 graph classification datasets [27] from three domains including seven molecules datasets (MUTAG, BZR, COX2, DHFR, PTC_MR, AIDS, NCI1), three proteins datasets (ENZYMES, DD, PROTEINS), and three social network datasets (COLLAB, IMDB-BINARY, IMDB-MULTI), each with a set of graphs. Node features are available in some datasets, and graph labels are either binary or multi-class. Details of the datasets are presented in the Appendix.

We design two settings that follow different data partitioning mechanisms. One setting is to randomly distribute graphs from a single dataset to a number of clients, with each client holding a distinct set of 30-50 graphs, among which 10% are held out for testing. In the other setting, we use multiple datasets either from a single domain or multiple domains. Each client holds a distinct set of 50 graphs from one dataset, among which 10% are held out for testing. In the first setting, we use NCI1, PROTEINS, and IMDB-BINARY from three domains and distribute them to 80, 30, 20 clients, respectively. In the second setting, we create three data groups including MOLECULES which consists of seven datasets from the molecules domain distributed into seven clients, BIOCHEM which adds three datasets from the proteins domain into MOLECULES and distributes them into ten clients, MIX which adds three datasets from the social domain into BIOCHEM and distributes them into 13 clients.

**Baselines** We use `self-train`[1] as the first baseline to test whether federated learning can bring improvements to each client through collaborative training. In `self-train`, each client firstly downloads the same randomly initialized model from the server and then trains locally without any communications. Then we implement two widely used FL baselines of `FedAvg` [25] and `FedProx` [23] which is a baseline that is able to deal with data and system heterogeneity. For the graph classification model, we use the same two-layer GIN [40], which represents the state-of-the-art GNN for graph-level tasks. We fix the GIN architecture and hyper-parameters through all baselines in order to control the experiments across different settings.

**Parameter settings** We use the two-layer GINs with hidden size of 64. We use a batch size of 32, and an Adam [19] optimizer with learning rate 0.001 and weight decay $5e^{-4}$. The $\mu$ for `FedProx` is set to 0.01. For all FL methods, the local epoch is set to 3. The two important hyper-parameters $\epsilon_1$ and $\epsilon_2$ as clustering criteria vary in different groups of data, which are set through offline training for about 50 rounds following [30]. We run all experiments for five random repetitions on a server with 8 24GB NVIDIA TITAN RTX GPUs.

[1] We do not have a global model as baseline as datasets are from various domains and their tasks are divergent.

## 6.2 Experimental results

**Federated graph classification within single datasets** Conceptually, clients in this setting are more homogeneous. As can be seen from the results in Table 3, our framework can obviously improve the performance of graph classification over local clients. For the NCI1 dataset distributed on 80 clients, GCFL and GCFL+ achieve $5.57\%$ and $6.26\%$ performance gains over `self-train`, and they help 17 more clients than `FedAvg` and 13 more clients than `FedProx`. For the PROTEINS dataset on the total 30 clients, the average performance gains over `self-train` are more significant, i.e., $12.88\%$ and $13.73\%$. Both GCFL and GCFL+ are able to improve all 30 clients. For IMDB-BINARY on 20 clients, both `FedAvg` and `FedProx` fail to improve the clients on average, and `FedAvg` can help only 5 clients. This is consistent with the results shown in Table 2 that the IMDB-BINARY dataset itself has relatively high structure and feature heterogeneity, which makes `FedAvg` ineffective. Our GCFL and GCFL+ frameworks can still improve the performance on IMDB-BINARY on average, and help nearly all clients (17&18 out of 20). These experimental results demonstrate that our frameworks are effective on the single-dataset multi-client FL setting.

**Federated graph classification across multiple datasets** According to our data analysis in Tables 1 and 2, clients in such a setting are more heterogenenous. As can be seen from the results in Table 4, our frameworks GCFL and GCFL+ can significantly improve the performance of clients with distinct datasets. We conduct experiments with multiple datasets in two settings: single domain (using the data group MOLECULES), and across domains (using the data groups BIOCHEM and MIX). The results show $6.34\% - 7.32\%$ improvements of our frameworks compared to `self-train`. A noticeable result is that our GCFL+ framework achieves a ratio of $100\%$ for all 3 data groups to improve all clients' performance. Additionally, the GCFL+ framework also outperforms GCFL. These results indicate that graphs across datasets or even across domains are able to help each other through proper FL, which is a surprising and interesting start point for further study.

Table 3: Performance on the single dataset multiple client setting. We present the average accuracy and minimum gain over `self-train` on all clients, as well as the ratio of clients which get improved.

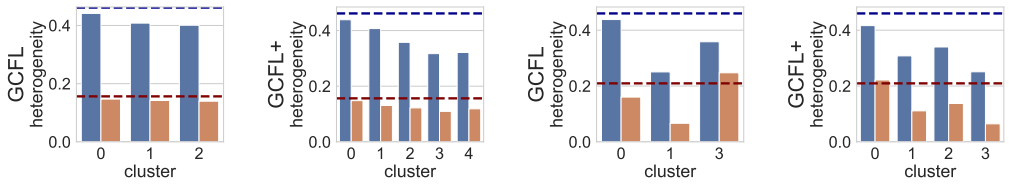
| Dataset (# clients) | NCI1 (80) | | | PROTEINS (30) | | | IMDB-BINARY (20) | | |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | average | min gain | ratio | average | min gain | ratio | average | min gain | ratio |
| `self-train` | 0.5874(±0.018) | — | — | 0.5979(±0.028) | — | — | 0.6178(±0.024) | — | — |
| FedAvg | 0.6013(±0.021) | -0.0423 | 56/80 | 0.6477(±0.027) | -0.0125 | 28/30 | 0.6122(±0.031) | -0.0300 | 5/20 |
| FedProx | 0.6019(±0.020) | -0.0443 | 60/80 | 0.6430(±0.023) | -0.0339 | 28/30 | 0.6117(±0.028) | -0.0380 | 10/20 |
| GCFL | 0.6201(±0.015) | -0.0326 | **73**/80 | 0.6749(±0.023) | 0.0268 | **30**/30 | 0.6345(±0.020) | -0.0100 | 17/20 |
| GCFL+ | **0.6242**(±0.016) | -0.0380 | **73**/80 | **0.6800**(±0.019) | 0.0143 | **30**/30 | **0.6374**(±0.023) | -0.0060 | **18**/20 |

Table 4: Performance on the multiple dataset multiple client setting. Metrics are the same as Table 4.

| Dataset (# domains) | MOLECULES (1) | | | BIOCHEM (2) | | | MIX (3) | | |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | average | min gain | ratio | average | min gain | ratio | average | min gain | ratio |
| `self-train` | 0.6992(±0.027) | — | — | 0.6405(±0.022) | — | — | 0.6136(±0.022) | — | — |
| FedAvg | 0.7133(±0.029) | -0.0211 | 4/7 | 0.6539(±0.030) | -0.0237 | 6/10 | 0.6307(±0.027) | -0.0322 | 9/13 |
| FedProx | 0.7082(±0.025) | -0.0468 | 3/7 | 0.6507(±0.032) | -0.0433 | 7/10 | 0.6237(±0.026) | -0.0383 | 8/13 |
| GCFL | 0.7356(±0.029) | -0.0010 | 6/7 | 0.6785(±0.018) | -0.0017 | 9/10 | 0.6526(±0.026) | -0.0068 | 12/13 |
| GCFL+ | **0.7435**(±0.024) | 0.0049 | **7**/7 | **0.6843**(±0.025) | 0.0010 | **10**/10 | **0.6585**(±0.026) | 0.0056 | **13**/13 |

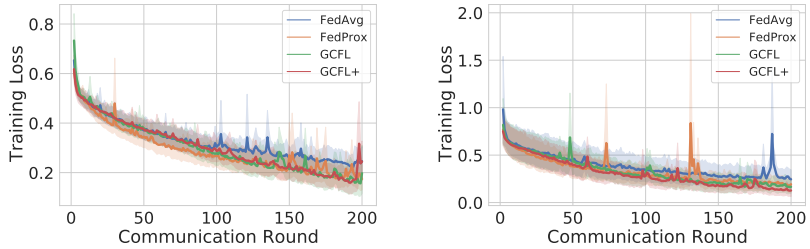
## 6.3 Structure and feature analysis in clusters

We conduct in-depth analysis to explore the clustering results of GCFL and GCFL+. As can be seen in Figure 2, after being clustered by GCFL and GCFL+, the overall structure and feature heterogeneity of clients' graphs within clusters are reduced significantly compared to the original values, especially for the multiple dataset setting (Figure 2c and 2d). For the one dataset setting (Figure 2a and 2b), since features all fall in the same space, pairs of clients tend to have more homogeneous features. Therefore, the feature heterogeneity only gets reduced slightly after clustering. Unlike feature heterogeneity, the structure heterogeneity within clusters decreases significantly. In the setting of multiple datasets, as shown in Figure 2c and 2d, both structure and feature heterogeneity decrease significantly, which is intuitive since datasets across domains usually tend to have higher heterogeneity, as discussed in 4.1. We also look into the clusters and find that datasets from the same domains are more likely to be clustered together, while datasets from different domains also constantly get clustered together and benefit each other. For example, the clustering of GCFL+ corresponding to Figure 2d groups two social networks COLLAB and IMDB-BINARY together with PROTEINS and also several molecules




(a) oneDS: PROTEINS (b) oneDS: PROTEINS (c) multiDS: MIX (d) multiDS: MIX

Figure 2: Structure (blue) and feature (red) heterogeneity within clusters found by GCFL and GCFL+. Dashed lines denote the heterogeneity over all clients before clustering.




(a) oneDS: PROTEINS (b) multiDS: MIX

Figure 3: Average with standard deviation of the training curves of all clients.

datasets, and there is also a cluster of NCI1, DD, and IMDB-MULTI which are molecules, proteins and social networks, respectively. These analysis manifests that domains of datasets can verify the sanity of clusters to some extent, but one cannot solely rely on such prior knowledge to determine the optimal clusters, which demonstrates the necessity of our frameworks with the ability of performance-driven dynamic clustering along the process of FL.

### 6.4 Convergence analysis

We visualize the testing loss with respect to the communication round to show the convergence of GCFL and GCFL+ compared with the standard federated learning baselines. Figure 3 shows the training curves on two settings, which illustrates that GCFL and GCFL+ achieves similar convergence rate as FedProx, which is the state-of-the-art FL framework dealing with non-IID Euclidean data. We also notice that both GCFL, GCFL+ and FedProx can converge to a lower loss compared with `FedAvg`, which corroborates our consideration of the non-IID problem in our setting.

### 6.5 More results in Appendix

In Table 3 and 4, we averaged the accuracy across all clients for presentation simplicity. To understand the detailed performance by clients and clusters, we present different Violin plots in the Appendix. Besides, we also show more results regarding various settings (overlapping clients, real vs. synthetic node features, standardized gradient-sequence matrix in GCFL+, etc) in the Appendix.

## 7 Conclusion

In this work, we propose a novel setting of cross-dataset and cross-domain federated graph classification. The techniques (GCF and GCF+) we develop allow multiple data owners holding structure and feature non-IID graphs to collaboratively train powerful graph classification neural networks without the need of direct data sharing. As the first trial, we focus on the effectiveness of FL in this setting and have not carefully studied other issues such as data privacy, although it is intuitive to preserve the privacy of clients by introducing an encryption mechanism (e.g. applying orthonormal transformations), and to prevent from adversarial scenarios by clustering out the malicious clients. Due to its evident motivations and proofs on the effective FL in a new setting, we believe this work can serve as a stepping stone for many interesting future studies.

# References


[1] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *ICDM*, 2005.

[2] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *IJCNN*, 2020.

[3] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

[4] Mingyang Chen, Wen Zhang, Zonggang Yuan, Yantao Jia, and Huajun Chen. Fede: Embedding knowledge graphs in federated setting, 2020.

[5] Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *NeurIPS*, 2021.

[6] P. Erdős and A Rényi. On random graphs. 1. *Publicationes Mathematicae.*, 6:290–297, 1959.

[7] Alireza Fallah, Aryan Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *NeurIPS*, 2020.

[8] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. In *NeurIPS*, 2020.

[9] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *NeurIPS*, 2020.

[10] E.N. Gilbert. Random graphs. *Annals of Mathematical Statistics.*, 30 (4):1141–1144, 1959.

[11] Li Huang, Yifeng Yin, Zeng Fu, Shifa Zhang, Hao Deng, and Dianbo Liu. Loadaboost: Loss-based adaboost federated machine learning on medical data. *PLoS ONE*, 15(4):e0230706, 2020.

[12] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *AAAI*, 2021.

[13] Sergey Ivanov and Evgeny Burnaev. Anonymous walk embeddings. In *ICML*, 2018.

[14] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[15] Meng Jiang, Taeho Jung, Ryan Karl, and Tong Zhao. Federated dynamic gnn with secure aggregation. *arXiv preprint arXiv:2009.07351*, 2020.

[16] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[17] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 2019.

[18] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *AISTATS*, 2020.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2017.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[21] Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173*, 2019.

[22] Jurij Leskovec, Deepayan Chakrabarti, Jon Kleinberg, and Christos Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *ECML-PKDD*, 2005.

[23] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 2020.

[24] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

[25] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017.

[26] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

[27] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020.

[28] Niels Lundtorp Olsen, Bo Markussen, and Lars Lau Rakêt. Simultaneous inference for misaligned multivariate functional data, 2017.

[29] Karl Pearson. Das fehlergesetz und seine verallgemeinerungen durch fechner und pearson. a rejoinder [the error law and its generalizations by fechner and pearson. a rejoinder]. *Biometrika*, 4 (1-2):169–212, 1905.

[30] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2020.

[31] Nino Shervashidze and Karsten M. Borgwardt. Fast subtree kernels on graphs. In *NIPS*, 2009.

[32] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *AISTATS*, 2009.

[33] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *JMLR*, 11:1201–1242, 2010.

[34] Binghui Wang, Ang Li, Hai Li, and Yiran Chen. Graphfl: A federated learning framework for semi-supervised node classification on graphs, 2020.

[35] Chunnan Wang, Bozhou Chen, Geng Li, and Hongzhi Wang. Fl-agcns: Federated learning framework for automatic graph convolutional network search. *arXiv preprint arXiv:2104.04141*, 2021.

[36] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

[37] Chuhan Wu, Fangzhao Wu, Yang Cao, Yongfeng Huang, and Xing Xie. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*, 2021.

[38] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.

[39] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 32(1):4–24, 2021.

[40] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[41] Pinar Yanardag and S.V.N. Vishwanathan. *Deep Graph Kernels*, page 1365–1374. Association for Computing Machinery, New York, NY, USA, 2015.

[42] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

[43] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *AAAI*, 2019.

[44] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[45] Jun Zhou, Chaochao Chen, Longfei Zheng, Huiwen Wu, Jia Wu, Xiaolin Zheng, Bingzhe Wu, Ziqi Liu, and Li Wang. Vertically federated graph neural network for privacy-preserving node classification. *arXiv preprint arXiv:2005.11903*, 2020.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Mainly see Sections 4-6
   (b) Did you describe the limitations of your work? [Yes] See Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] See Section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.4
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 6.1 and the supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For example see Tables 3 and 4
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Wherever we mention them for the first time
   (b) Did you mention the license of the assets? [N/A] They are public
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our models and code in the supplemental material
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]