Technical Appendix for Modelling with Homophily Driven Heterogeneous Data in Gossip Learning

1 Characterization of SDAH

To make the theory tractable, let us consider a binary classification problem, i.e., $|C| = 2$. So, instead of a dirichlet distribution, it suffices to consider a beta distribution with probability density function as $P(z) = \frac{1}{B(\gamma/2,\gamma/2)}(z(1-z))^{(\gamma/2-1)}$, where $B(.)$ denotes the beta function [1].

Next let us analyse SDAH graphs using both theory and simulations.

Theorem 1. *1 If each node i samples $Z \sim beta(\gamma,\gamma)$ with $\gamma < 1$ then the probability, p that none of the n nodes will have Z value from the interval $[1/2 - k\epsilon, 1/2 + k\epsilon]$ for $0 < \epsilon < 1$ and $1 < k < 1/\epsilon$ is:*

$$\left(1 - \frac{(1/4 - k^2\epsilon^2)^{\gamma-1}}{B(\gamma,\gamma)}\right)^{2nk} \leq p \leq exp\left(-\frac{2nk}{4^{\gamma-1}B(\gamma,\gamma)}\right)$$

Proof. The probability density function for dirichlet distribution is $P(Z) \propto \prod_{i=1}^{|C|} Z_i^{\frac{\gamma}{|C|}-1}$, $\sum_i Z_i = 1$).

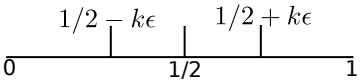


Figure 1: Setting for Theorem 1.

Consider the feature space as shown in Figure 1. From the PDF of beta distribution we have $P(Z = 1/2) = \frac{1}{B(\gamma,\gamma)4^{\gamma-1}}$ and $P(Z = 1/2 - k\epsilon) = \frac{1}{B(\gamma,\gamma)}(1/4 - k^2\epsilon^2)^{\gamma-1}$.

From independence of sampling and symmetry of the problem we have, $P(Z \notin [1/2 - k\epsilon, 1/2 + k\epsilon]) = P(Z \notin [1/2 - k\epsilon, 1/2])P(X \notin (1/2, 1/2 + k\epsilon]) = P(Z \notin [1/2 - k\epsilon, 1/2])^2$.

Upper bound. As $\gamma < 1$, we have $P(Z \notin [1/2-k\epsilon, 1/2]) = P(Z \neq 1/2-k\epsilon)P(Z \neq 1/2-(k-1)\epsilon)\cdots P(Z \neq 1/2) \leq P(Z \neq 1/2)^k = (1 - \frac{1}{B(\gamma,\gamma)4^{\gamma-1}})^k \leq \exp(\frac{-k}{B(\gamma,\gamma)4^{\gamma-1}})$

Now, $P(Z \notin [1/2 - k\epsilon, 1/2 + k\epsilon]) \leq \exp(\frac{-2k}{B(\gamma,\gamma)4^{\gamma-1}})$ and the probability that none of the n samples land in this interval is $\exp(\frac{-2k}{B(\gamma,\gamma)4^{\gamma-1}})$.

Lower bound. As $\gamma < 1$, we have $P(Z \notin [1/2-k\epsilon, 1/2]) = P(Z \neq 1/2-k\epsilon)P(Z \neq 1/2-(k-1)\epsilon)\cdots P(Z \neq 1/2) \geq P(Z \neq 1/2 - k\epsilon)^k = (1 - \frac{1}{B(\gamma,\gamma)}(1/4 - k^2\epsilon^2)^{\gamma-1})$

Computing the probability for both the sides and for n draws produces the statement.

□

The above theorem shows that with increasing n and $\gamma < 1$ with high probability the nodes will be well separated in the Z space. The statement can be trivially extended to the case of concentration as $\gamma/2$,

[1] https://en.wikipedia.org/wiki/Beta_function

replacing the appropriate values. Now let us construct the communication network on this heterogeneous data distribution. A node i connects to another node j with probability

$$P_{ij} = \frac{1}{1 + \left(\frac{\max(k_{ij}, k_{ji})}{b}\right)^{\xi}} \tag{1}$$

Below, we characterize how clustered the generated graph is using edge expansion Kempe (2011). For a set of nodes S_i it is defined as $\rho(S_i) = \frac{|L(S_i, \overline{S_i}|)}{min(|S_i|, |\overline{S_i}|)}$ where $\overline{S_i}$ denotes all nodes not in S_i and $L(S_i, \overline{S_i})$ denotes the set of edges connecting S_i and $\overline{S_i}$. A lower ρ denotes sparser inter-cluster connection.

Theorem 2. *2 Given two well separated clusters of point clouds (e.g., the ones in Theorem* 1*), SDAH would produce graphs (with n nodes) with the following expected edge expansion numbers for a cluster S_i with m nodes.*

$$\mathbb{E}[\rho(S_i)] \leq \begin{cases} m/2, \ \textit{for } \xi = 0 \\ b \ln\left(\frac{n+b}{m+b}\right) \ \textit{for } \xi = 1 \\ b \arctan\left(\frac{b(n-m)}{b^2+(m+b)(n+b)}\right) \ \textit{for } \xi = 2 \end{cases}$$

Note that the growth rate of $\arctan(x)$ is much less than $\ln(x)$ for $x > 0$, for example, $\arctan(x) \leq \pi/2$ whereas $\ln x$ is monotonically increasing function without such a upper bound. Thus ρ decreases when ξ increases.

Proof. **When $\xi = 0$.** In this case the probability of any pair nodes to be connected is $1/(1 + k^0) = 1/2$. Thus, the expected number of edges connecting the two clusters is $(n-m)m/2$. Now, considering $m \approx n-m$, we get the expected edge expansion as $\frac{m^2}{2m} = m/2$.

When $\xi = 1$. Here for a node i, $P_{ij} = \frac{1}{1+d/b}$ where $d \in [1, n]$. The expected number of edges from a node in C_1 to the other cluster is $= \frac{1}{1+(m+1)/b} + \frac{1}{1+(m+2)/b} + \cdots + \frac{1}{1+(n)/b} = \sum_{i=1}^{n} \frac{1}{1+i/b} - \sum_{i=1}^{m} \frac{1}{1+i/b} = b\ln(n/b) - b\ln(m/b) = b\ln(\frac{n+b}{m+b})$.

Note that this differs from the P_{ij} that needs to compute the maximum of k_{ij} and k_{ji}. However, the probability above will always be larger than the actual probability and thus the expected value we compute here is a conservative upper bound to the actual expectation.

Thus the expected number of inter-cluster edges is $mb\ln(\frac{n+b}{m+b})$. Further, the expected edge expansion is $mb\ln(\frac{n+b}{m+b}/m)$.

When $\xi = 2$. Here for a node i, $P_{ij} = \frac{1}{1+(d/b)^2}$ where $d \in [1, n]$. The expected number of edges from a node in C_1 to the other cluster is $= \frac{1}{1+((m+1)/b)^2} + \frac{1}{1+((m+2)/b)^2} + \cdots + \frac{1}{1+(n/b)^2} = \sum_{i=1}^{n} \frac{1}{1+(i/b)^2} - \sum_{i=1}^{m} \frac{1}{1+(i/b)^2} = b\arctan(n/b) - b\arctan(m/b)$. Utilizing the relation $\arctan(\alpha) - \arctan(\beta) = \arctan(\frac{\alpha-\beta}{1+\alpha\beta})$, we get the statement.

□

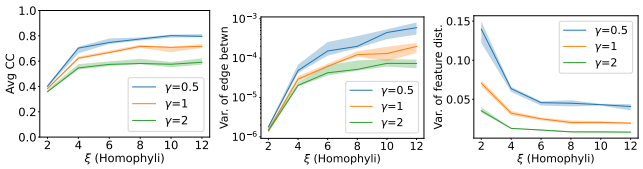
Empirical characterization of SDAH. Figure 2 presents how the community structure and data distribution across nodes for the SDAH graphs change with γ and ξ. All simulations use $b = 10$ and we analyze 10 randomly generated graphs of 100 nodes to show median, 25, and 75 percentiles.

Clustering Coefficient (CC) Saramäki et al. (2007) of a node u is the fraction of possible triangles containing u – it increases for more connected graph, for example it is 1 for a complete graph. A typical social network has high CC enforcing the idea that people with common friends tend to connect Kempe (2011). SDAH graphs have high CC (Figure 2(a)) for a wide range of values for ξ and γ confirming that our framework captures the realistic characteristics of social networks. CC increases with increasing ξ (communities become more densely connected) and decreasing γ (as the data distribution becomes more heterogeneous, the nodes become more clustered in the feature space). Moreover, in our experiments we see that SDAH graphs have small diameters similar to real-world social networks (More results in the technical appendix).

Betweenness centrality Brandes (2008) of an edge e is the fraction of all-pairs shortest paths that pass through e. Edge betweenness centrality of the bridge edges (connecting different clusters) are higher than

the intra-cluster edges. These values increase when a graph has a small number of bridges. Thus an increase in the variance of edge betweenness values reflect sparser inter-cluster connections (Figure 2(b)). Note that this measure is independent of any community detection method and thus is a general characterization of the topological structure.

Next we measure the Euclidean distance between the feature vectors (Z) of connected nodes and report their variance in Figure 2(c). It confirms that with increasing homophily, it is unlikely to connect two nodes with different features.



(a) (b) (c)

Figure 2: Shows how average clustering coefficient (CC) **(a)**, variance of the edge betweenness centrality **(b)**, and the difference in the data distribution over a pair of connected nodes **(c)** change with varying ξ and γ.

2 Experimental details

Datasets. Both MNIST and Fashion MNIST have 28×28 monochrome images from 10 classes. These images are of handwritten digits and fashion apparel respectively. Cifar-10 has colored 32×32 images from 10 classes. UCI-HAR dataset contains accelerometer signals while the participants performed one of the 6 activities.

Face to face contact graph of primary School students Génois and Barrat (2018). Here two students connect if they came face to face (tracked using RFID tags) at least 80 times over the course of the dataset and then we consider an edge to be there with probability 0.3 to account for the fact that not all proximity events result in interaction. The students from the same section naturally create clusters. We assume that the students in the same section have access to similar classes in their local data. Thus, we randomly assign 4 classes from MNIST to each cluster. Each node samples 100 training data points from 3 classes from its cluster-classes.

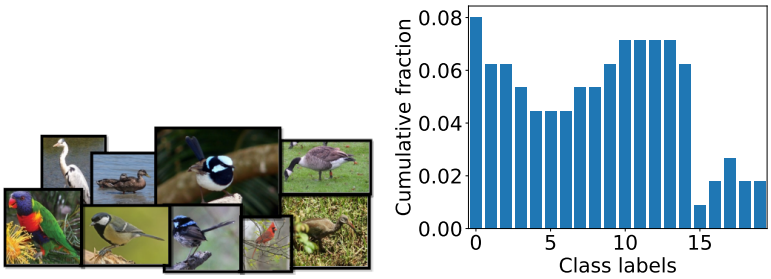
Face to face contact graph of village residents Ozella et al. (2021). Here the nodes are the residents of a village in rural Malawi. Two nodes connect when they met face to face during the first 2 days of the dataset. The face to face contact is tracked using a custom sensor worn by all the participants. We subsample 19 nodes by traversing in a breadth-first way starting from a random node in the network.

The network has natural clusters according to family memberships. We assume that the members of the same family have access to the same classes. Like the data distribution for the primary school, we sample training examples at each node.

Bird image classification. Here we build a dataset from iNaturalists. While there are many publicly available datasets derived from the same source [2] unfortunately none of them contains a communication network. Thus we use the following mechanism to build a communication network. We choose 20 popular bird species from four continents (5 from each) North America, Africa, Europe, and Australia. Then identified the contributors who have uploaded many pictures of these species. We use the locations of the contributed photos as a proxy for visited locations. The similarity between two individuals is measured using Jaccard similarity [3] of the visited states. We connect two people in the graph if the similarity exceeds 0.05.

[2] https://www.kaggle.com/c/inaturalist-2021

[3] Jaccard similarity of two sets A and B is $\frac{|A \cap B|}{|A \cup B|}$



(a) (b)

Figure 3: Sample bird images from the dataset.

The nodes create clusters according to their home countries. Home country of a person is computed as the country where s/he has the most number of photos taken. The clustering is natural as many people travel nearby and a very few go far, for example abroad. For each individual in this network,

We select the four most contributed classes at each node. This is to alleviate the heterogeneity due to the long tail distribution of the samples at a node. The bird images are sampled from a Kaggle bird dataset [4] and iNaturalist. At each node, we sample 50 training examples for each class. A few example pictures from the dataset are shown in Figure 3. There are different types of images, for example, some containing multiple birds, some with the bird in the focus whereas in some photos the bird is a bit obscured in the background. All the images are resized to 128×128 as part of preprocessing.

The aggregate distribution of classes across nodes is not balanced (Figure 3(b)). All the rare classes correspond to the species from Africa.

Models and optimizer. For MNIST, we use a model with 2 convolution layers followed by two fully connected layers. For Cifar10 we use a model with three blocks followed by two fully connected layers. Each block contains two convolution layers followed by a maxpool layer and a batch normalization layer. Fashion MNIST uses a model with two blocks followed by three fully connected layers. Each block contains a convolution layer followed by a maxpool layer and a batch normalization layer. UCI-HAR dataset uses a model with two 1-D convolution layers followed by a maxpool layer and two fully connected layers. These models do not attain state of the art performance, but suffices our purpose.

All the experiments use SGD optimizer with batch size 32, learning rate of 0.01, and momentum of 0.9.

Other configuration. All experiments use 10 epochs in between the rounds except the one with CIFAR-10 which uses 1 local epoch. All experiments were performed on a single GPU machine and the code uses Pytorch library for implementation. The experiments use a single threaded process.

3 Evaluating on more synthetic graphs.

We synthetically generate communication graphs with varying separation of two clusters. Classes are distributed similarly as the toy dataset. We create two Erdos-Renyi Gilbert (1959) clusters with size 20 and connection probability 0.3. We vary the probability of connecting an inter-cluster node pair as $0.1, 0.05, 0.01,$ and 0.005 which produce modularity values Clauset et al. (2004) as $0.18, 0.36, 0.39,$ and 0.46 (higher values denote lower number of bridges). Figure 6 shows that our algorithm again converges faster than equal weighting.

4 Contact graphs

Here we show a few contact graphs we used in our experiments in Figure 5.

[4] https://www.kaggle.com/datasets/gpiosenka/100-bird-species

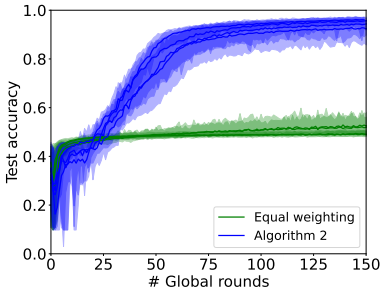


Figure 4: Our algorithm achieves much faster convergence speed than the baseline in a variety of synthetic settings.

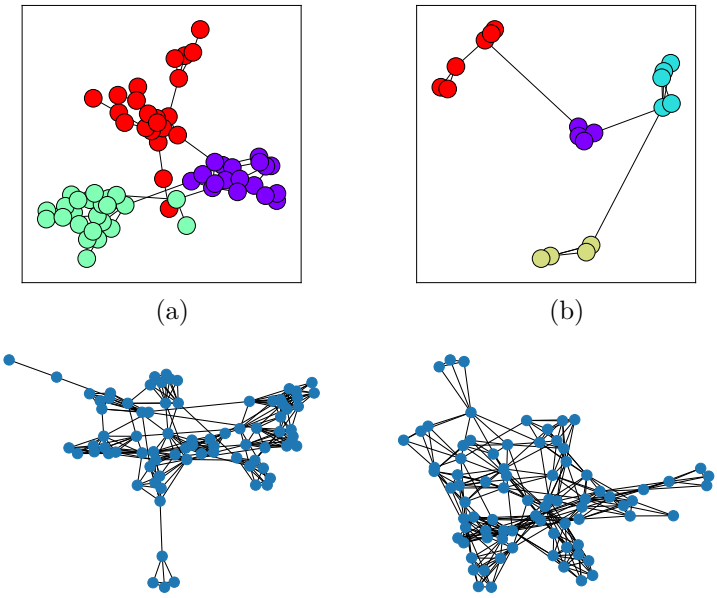


Figure 5: **(a)** Primary school student contact graph. Different colors represent students of different classes. **(b)** Contact graph for village residences and the color represents different families. **(c)** SDAH graph with $\gamma = 1$, $\xi = 8$, and $b = 10$. **(d)** SDAH graph with $\gamma = 2$, $\xi = 8$, and $b = 10$. We can visibly see that the clustered structure changes with γ.

5 Characterizing SDAH

In Figure 6 we show that the SDAH graphs have small diameter as real social networks.

References

Ulrik Brandes. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social networks* 30, 2 (2008), 136–145.

Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

Mathieu Génois and Alain Barrat. 2018. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science* 7, 1 (2018), 11.

Edgar N Gilbert. 1959. Random graphs. *The Annals of Mathematical Statistics* 30, 4 (1959), 1141–1144.

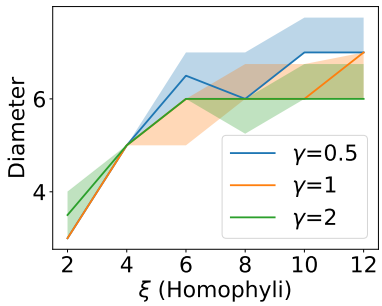


Figure 6: Diameter of SDAH graph with 100 nodes.

David Kempe. 2011. Structure and dynamics of information in networks. *Lecture Notes* (2011).

Laura Ozella, Daniela Paolotti, Guilherme Lichand, Jorge P Rodríguez, Simon Haenni, John Phuka, Onicio B Leal-Neto, and Ciro Cattuto. 2021. Using wearable proximity sensors to characterize social contact patterns in a village of rural Malawi. *EPJ Data Science* 10, 1 (2021), 46.

Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. 2007. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E* 75, 2 (2007), 027105.