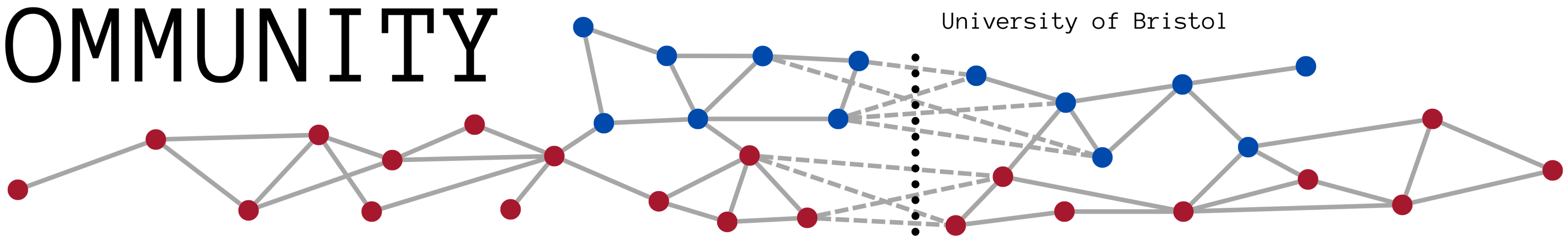
University of BRISTOL

arXiv: Website:

A FRAMEWORK FOR EXPLORING FEDERATED COMMUNITY DETECTION

WILLIAM LEENEY AND RYAN MCCONVILLE
{will.leeney, ryan.mcconville}@bristol.ac.uk
School of Engineering Mathematics and Technology
University of Bristol



Federated Learning is machine learning between a network of clients whilst maintaining data residency and/or privacy constraints. **Community Detection** is the unsupervised discovery of clusters of nodes within graph-structured data. **Federated Community Detection** is challenging due to the complexity induced due to missing connectivity information between privately held graphs.

01 REAL-WORLD APPLICATIONS

- Banks: collaboration on anti-fraud measures that can't share client information -> each client owns a sub-section of a global graph.
- Social Media: preventing fake news propagation between platforms -> data is heterogeneous (comment != tweet)
- Pharmaceuticals: molecular copyright ownership -> expensive proprietary feature space information restriction
- Hospitals: can't share user identities -> restrictions on connectivity space

02 PROBLEM STATEMENT

C : Number of Clients

$G = [G_1, G_2, \dots, G_C]$: Partitioned Graph

$= [\{X_1, A_1\}, \{X_2, A_2\}, \dots, \{X_C, A_C\}]$

N_c, M_c : Each own nodes and edges

$Y_C \in \{1, 2, \dots, k\}$: Output Community Assignments

Constraints:
- Clients don't share data
- Loss function doesn't use labels

03 PROPOSED SOLUTION

(performed privately by each client)

$A' = D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}}$: normalise adjacency using degree vector: $d_u = \sum_{v=1}^{N} A_{uv}$

$X_{\beta_u=} \{x_v \in v \,|a'_{uv} = 1\}$: feature neighbourhood

$h_u = \sigma(b_u + \Sigma_{v\in\beta_u} a'_{uv}\psi(x_v))$: node-wise message passing

$F(X) = [h_1, h_2, \dots, h_N]^T$: GNN function

$Y = \xi(F(X))$: forward pass $Y' \in \{0,1\}^{N\times k}$: relax to one-hot assignments

$L\,(G, w) = -\frac{1}{2M} Tr\Big(Y'^T AY' - Y'd^T dY'\Big) + \lambda_r \left(\frac{\sqrt{k}}{N} \| \sum_u Y'_u \|_F - 1\right)$ [1]

$w_c\,(r+1) \leftarrow w_c\,(r) - \eta \bigtriangledown L$: update weights with DMoN loss

$w' \leftarrow \sum_{c}^{C=1} \frac{N_c}{N} w_c$: aggregate weights at server

: repeat for r local rounds and r_g aggregation rounds

σ : SELU
$\| \, \|_F$: Frobius Norm
ψ, ξ : neural networks
η : learning rate

04 EXPERIMENT DETAILS

- Local rounds: r = 5
- Communication rounds: r_g = 250
- Cluster size regularisation: λ_r = 1.0
- Adam optimiser using a learning rate: η = 0.001
- Dimensionality of hidden space features: 64
- Training/Validation/Test splits: 0.7/0.1/0.2
- W Randomness Coefficient quantifies the consistency of algorithm ranking comparisons over random seeds.

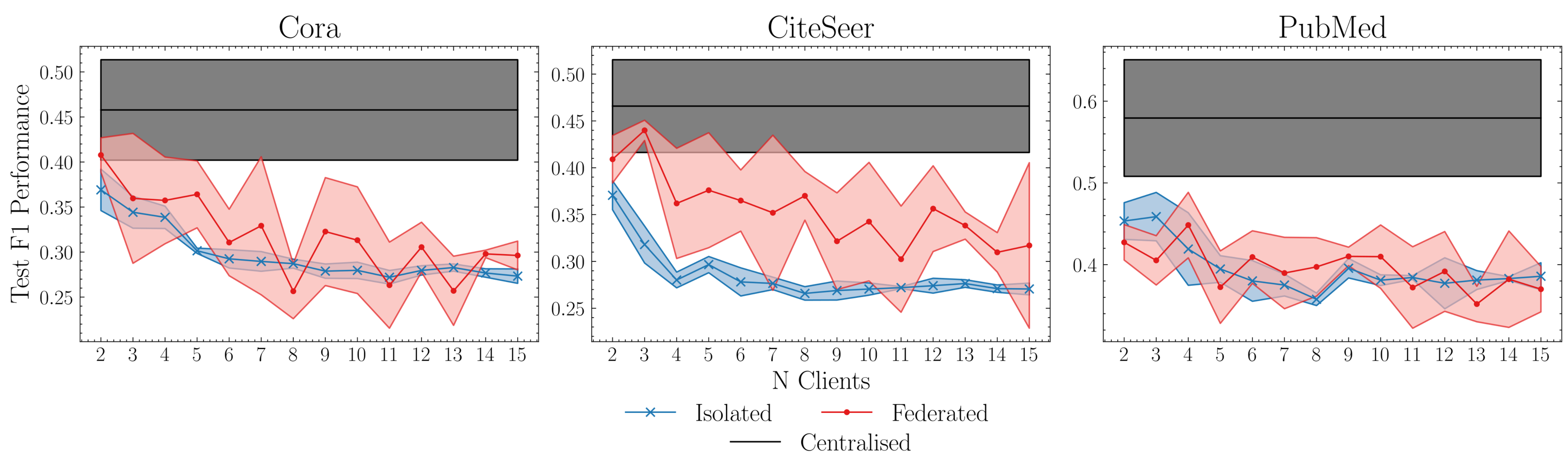
$$W = 1 - \frac{1}{T}\sum_{t\in T} \frac{12S}{n_s\,(n_{a^3} - n_a)} \quad [2]$$

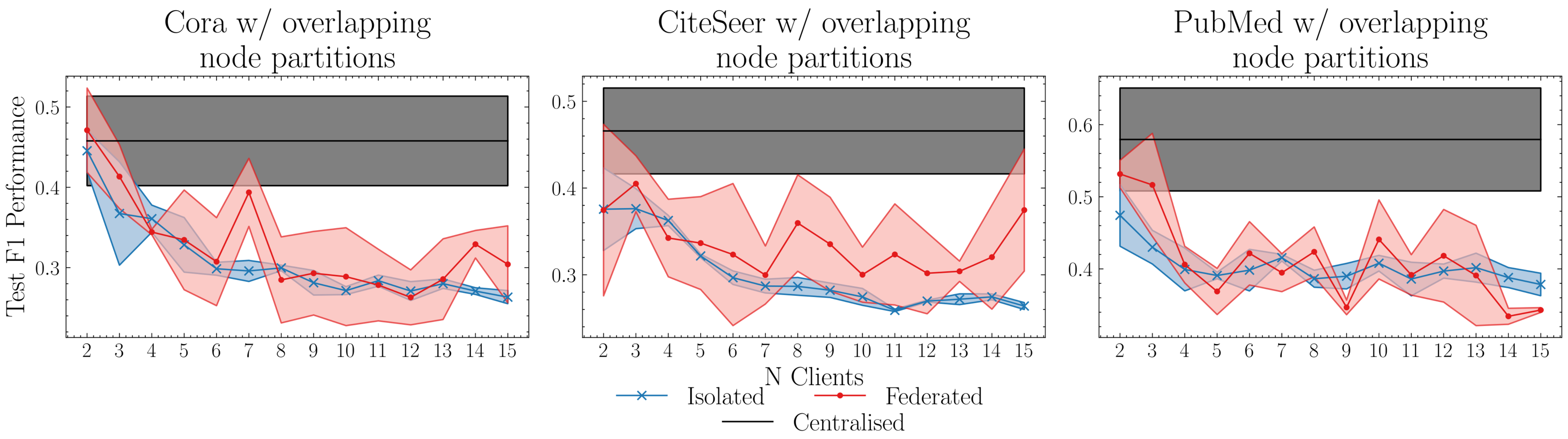
n_s :(3) number of seeds
n_a :number of algorithms
S :sum of squared deviations
T :set of benchmarking tests

Experiment 1 (W Randomness Coefficient = 0.212): No overlapping/shared nodes between clients in dataset partitioning.
Experiment 2 (W Randomness Coefficient = 0.259): Nodes may overlap between clients.

Datasets	Cora	CiteSeer	PubMed
n nodes	2708	3327	19717
n features	1433	3703	500
n edges	10556	9104	88648

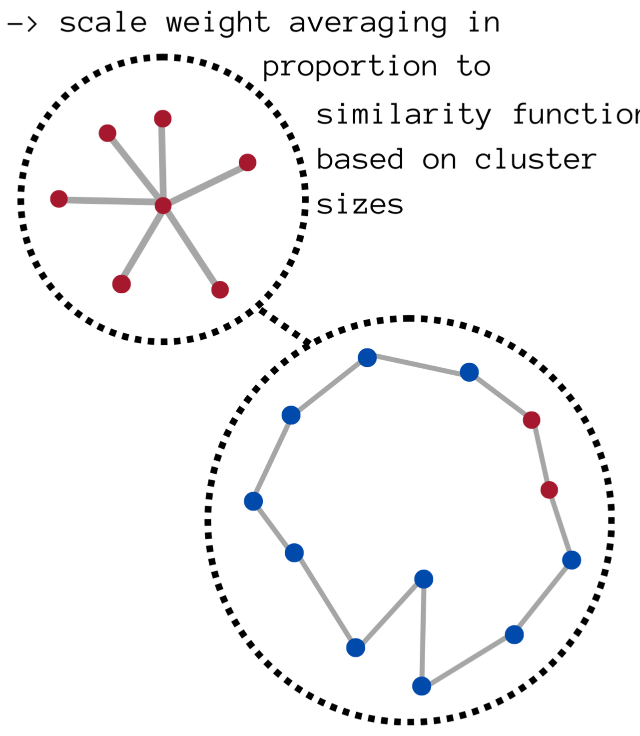
05 RESULTS



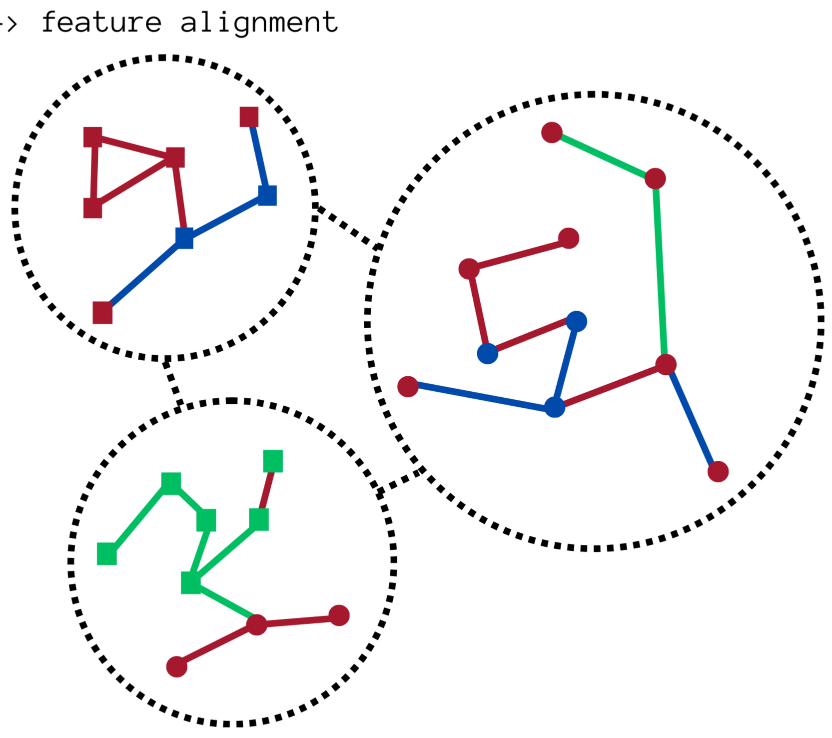


06 FUTURE DIRECTIONS

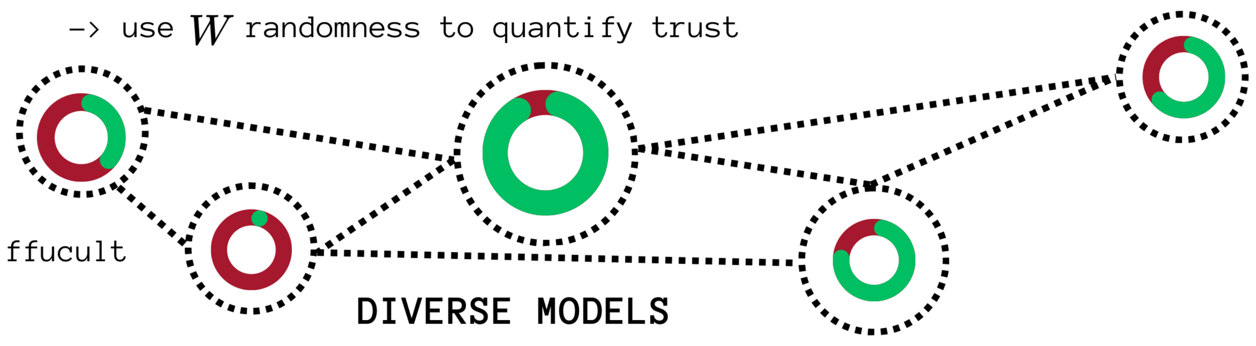
NON-IID DATA
-- different cluster sizes
-> scale weight averaging in proportion to similarity function based on cluster sizes



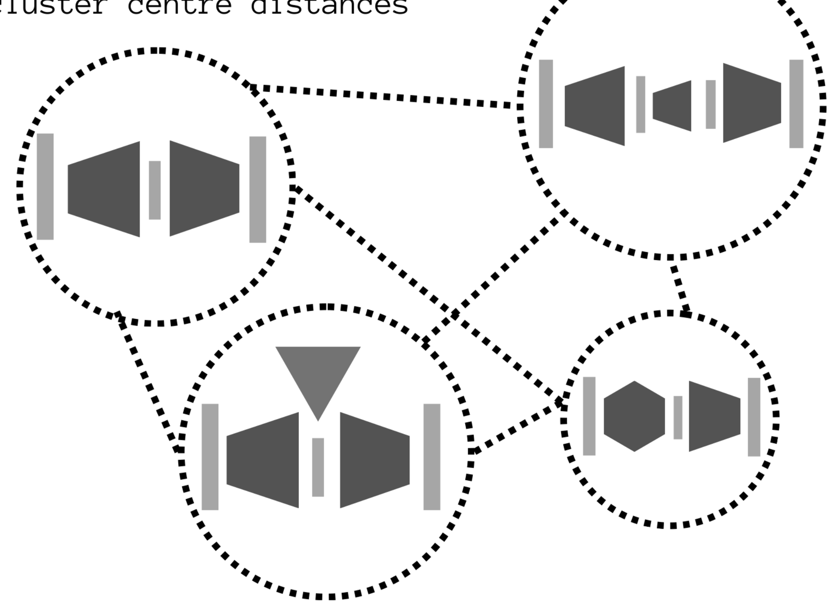
HETEROGENEOUS PARITIONS
-- aggregation of different modalities diffucult
-> feature alignment



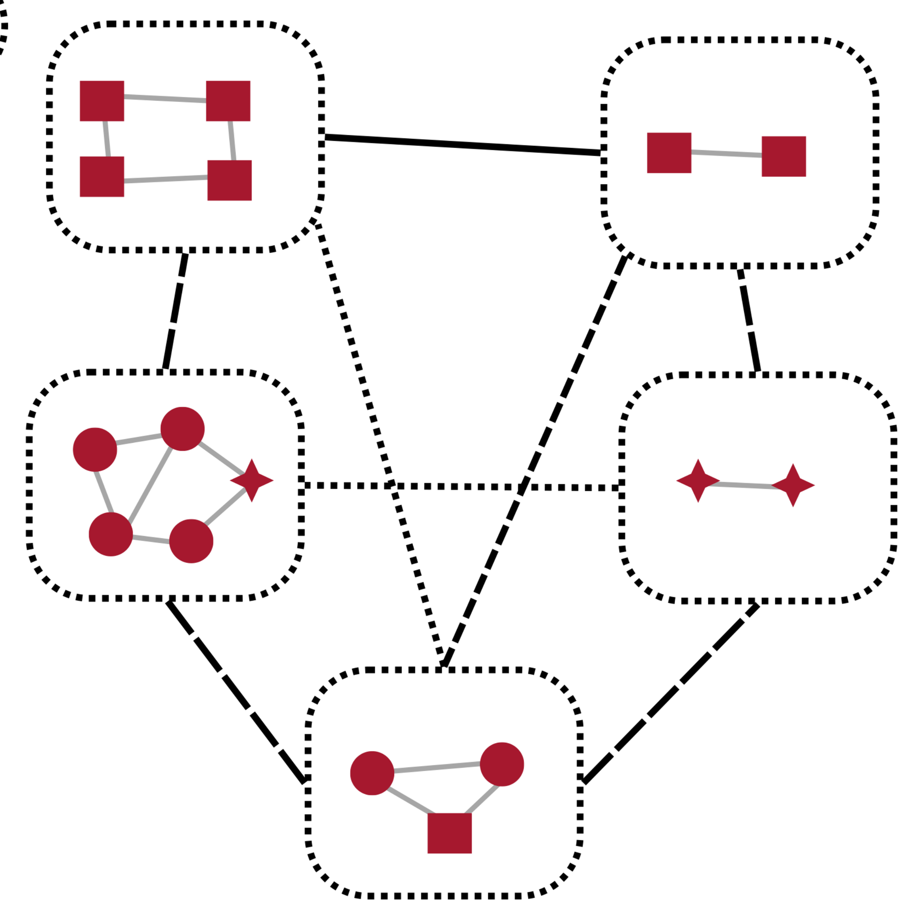
MALICIOUS ATTACK VULNERABILITY
-- some clients will act dishonestly + share bad weights
-> use W randomness to quantify trust



DIVERSE MODELS
-- cannot average weights due to size mismatch
-> minimise cluster centre distances



EFFICIENT COMMUNICATION
-- equal sharing of information is inefficient
-> share with similar clients



[1] Tsitsulin, Anton, John Palowitch, Bryan Perozzi, and Emmanuel Müller. "Graph clustering with graph neural networks." Journal of Machine Learning Research 24, no. 127 (2023): 1-21.
[2] William Leeney and Ryan McConville. Uncertainty in GNN Learning Evaluations: The Importance of a Consistent Benchmark for Community Detection. Twelfth International Conference on Complex Networks \& Their Applications (2023).