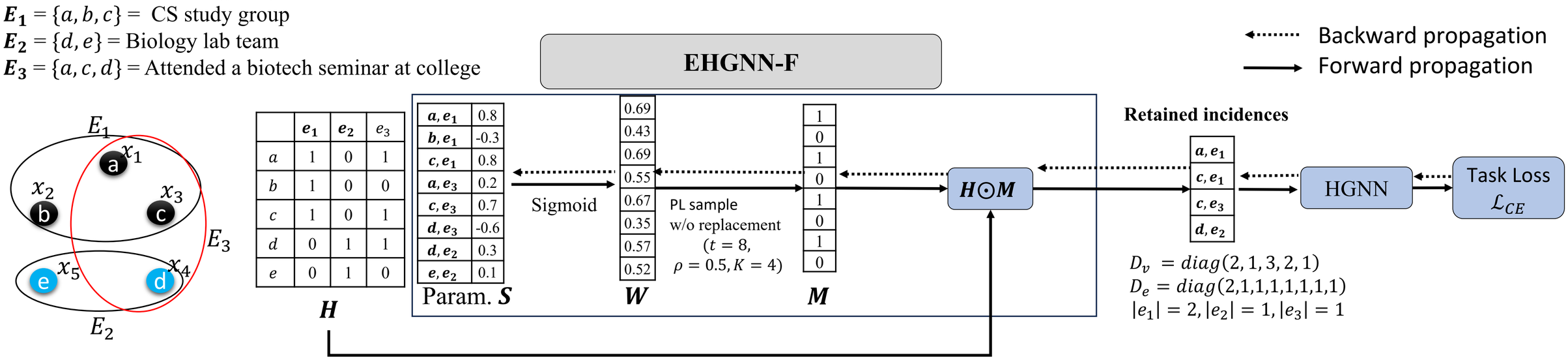
Supervised Incidence Sparsification of Hypergraphs under Budget Constraints

Naheed Anjum Arafat Danda B. Rawat
Naheedanjum.Arafat@howard.edu danda.rawat@howard.edu
DoW Center of Excellence in AI/ML (CoE-AIML)
Howard University, Washington, DC, USA

☞ Contribution: Learning to retain K incidence (node-hyperedge connections)

a, b, c = Students in CS major
d, e = Students in Biology major
$\boldsymbol{E_1} = \{a, b, c\}$ = CS study group
$\boldsymbol{E_2} = \{d, e\}$ = Biology lab team
$\boldsymbol{E_3} = \{a, c, d\}$ = Attended a biotech seminar at college



EHGNN-F in action: During training, sample exactly K incidences; at inference, retain the top-K.

1. Problem Statement

Motivation: Node-hyperedge connections can be redundant or poorly aligned with the prediction task.

Question: Given a budget, can task supervision select better connections than uniform random selection?

Formulation.
Budget-constrained, task-aware sparsification: Given incidence matrix $\boldsymbol{H}$, node features $\boldsymbol{X}$, and training labels, learn a binary mask that retains exactly K observed incidences for prediction:

$$\widetilde{\boldsymbol{H}} = \boldsymbol{H} \odot \hat{\boldsymbol{M}}, \qquad \|\hat{\boldsymbol{M}}\|_0 = K = \lfloor \rho t \rfloor.$$

t=number of incidences; ρ= retention ratio. Hyperedges are allowed to be partially retained. Self-loops are excluded from budget.

2. EHGNN-F: learn. sample. then rank

EHGNN-F is the incidence-level realization of **EdgeMask-HGNN**.

1. **Score.** Assign a learnable logit s_i to every observed incidence and convert it to a positive weight $w_i = \sigma(s_i)$.
2. **Sample during training.** Draw exactly K distinct incidences without replacement using Plackett–Luce sampling. Train the HGNN on this hard mask.
3. **Optimize jointly.** Backpropagate the task loss through a selected-only straight-through estimator (STE).

$$m_i = \hat{m}_i + (w_i - \text{stopgrad}(w_i))\hat{m}_i,$$

$$\underbrace{m_i = \hat{m}_i}_{\text{binary forward pass}}, \qquad \underbrace{\frac{\partial m_i}{\partial s_i} = \hat{m}_i w_i (1 - w_i)}_{\text{selected scores receive gradients}}.$$

stopgrad preserves the forward value and blocks its derivative. The STE is a biased gradient surrogate.

At inference: retain the deterministic **top-K** scores from the checkpoint with the lowest validation loss. **Both training and inference satisfy the exact budget.**

3. Theory and scope

Local sampler sensitivity. For $K < t$, let P_τ and $P_{\tau+1}$ be the size-K sampling distributions before and after a logit update.

$$\mathcal{W}_\triangle(P_\tau, P_{\tau+1}) \le \min\left\{2K,\ \frac{K}{2\alpha_\tau} \log \frac{t}{t-K} \|\boldsymbol{s}^{(\tau+1)} - \boldsymbol{s}^{(\tau)}\|_1\right\}.$$

$\mathcal{W}_\triangle$ is the expected number of differing incidences under an optimal coupling; $\alpha_\tau > 0$ is the minimum sampling weight across the two iterates.

Small score updates limit distributional change when the weights stay away from zero. This is a **conditional local bound**; it does not establish training convergence or stability of independently drawn masks.

Practical limits. The budget K is user-specified; the STE does not guarantee a globally optimal subset. The six-dataset mean gain over Random-Fixed is not statistically distinguishable from zero at any tested budget.

4. Does learned selection beat random?

Six datasets, 10 paired seeds, one 50/25/25% train/validation/test split per dataset. **Random-Fixed** samples one uniform exact-K mask and keeps it throughout training and evaluation.

Accuracy gain over Random-Fixed (percentage points)

Retention	10%	20%	30%	50%
Actor	**+1.02***	**+1.95***	**+2.41***	**+2.14***
Yelp	**+0.87***	**+1.63***	**+1.71***	**+1.45***
Walmart	−0.31†	−0.75†	−0.70†	**+1.07***
Twitch	+0.03	+0.03	+0.16	+0.18
Pokec	+0.01	−0.04	+0.06	+0.15
DBLP-CA	−0.21	−0.70†	−1.20†	−2.20†

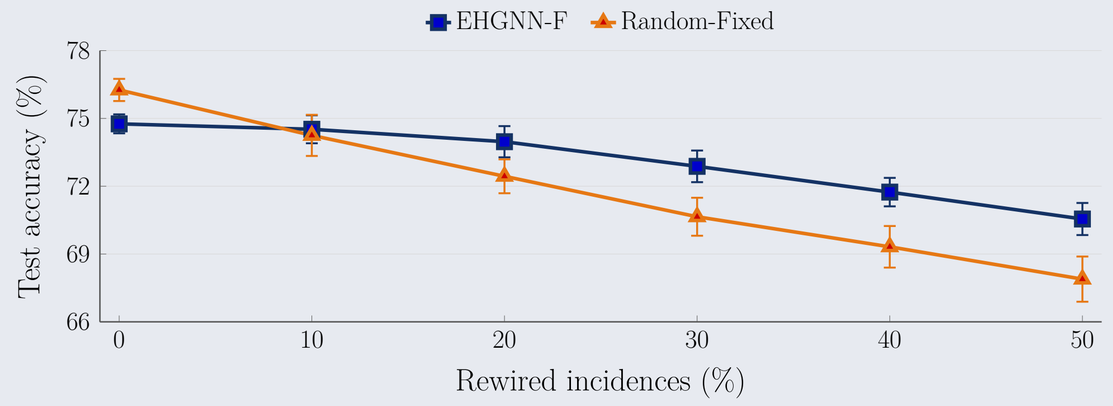
Differences of the rounded means in paper Table 2. * significant gain; † significant loss. Exact paired Wilcoxon tests, Holm correction over 24 comparisons, $p_H \le 0.05$.

Actor and Yelp: gains at every budget. These gains also hold against uniform training-time resampling, so resampling alone does not explain them.

Context matters: a feature-only MLP still outperforms **EHGNN-F** on Actor and Pokec. On Yelp, **EHGNN-F** beats the MLP but remains below Full's mean accuracy.

5. When incidences become misleading

Cora: rewire 0–50% of incidences to encourage class-mixed hyperedges; retrain at **50% retention**. Rewiring identities are hidden from all methods.



Means ± standard deviations over 10 paired seeds; concentrated rewiring ($\alpha = 1$). Data: camera-ready corruption study.

A reversal: worse on clean Cora, but **+1.54 to +2.66 points** over Random-Fixed at 20–50% corruption (Holm-significant).

What gets removed? EHGNN-F retains only **25–37%** of rewired incidences, versus about **50%** for Random-Fixed.

6. The computational trade-off

Yelp at 50% retention (three timing seeds):

Method	Fit/epoch (s)	Total (s)
Full	0.6693	669.27
EHGNN-F	0.5148	274.53
Random-Fixed	0.3139	287.37

23% lower fit/epoch time than Full, with extra scoring and sampling cost relative to fixed-mask sparsifiers.

Fit/epoch includes training and final evaluation divided by epochs run. Total also includes preprocessing and checkpoint selection; stopping times vary.

Acknowledgments. Supported in part by CoE-AIML at Howard University, U.S. Army Research Laboratory Contract W911NF-20-2-0277, and NSF grant #2602875 (CCP/CSE – Core Programs). .