Adversarial Training and Transfer on top of EGI

Abdullah Garra and **Mohammed Qaiss**
Tel Aviv University
{abdullahg,mohammedq}@mail.tau.ac.il

Abstract

We revisit Ego-Graph Infomax (EGI) and examine whether small adversarial edge edits improve cross-graph transfer. Using a lightweight PGD-style pretraining, we observe at most small and inconsistent gains. To understand this, we diagnose what the embeddings encode and find that degree is linearly recoverable from embeddings, and once we remove that signal, accuracy drops a lot. Moreover, a trivial no-encoder baseline challenges EGI on this benchmark. These findings try to explain why local, (almost) degree-preserving edits have limited effect. We also examine source selection and see that alternative sources—including randomized or degree-preserving variants—can rival or surpass those deemed best by EGI's transferability predictor (Δ_D). Overall, our results suggest that degree accounts for much of the apparent transfer on this benchmark and call into question the usefulness of Δ_D as a stand-alone proxy.

Code: github.com/abdullahgarra/advEGI

1 Introduction

Graph representation learning aims to learn node embeddings that transfer across domains. *Ego-Graph Infomax* - EGI (egi, 2021), models graphs as distributions over k-hop ego-structures and maximizes mutual information between local encodings and a global summary, with a theory linking transfer to a structural discrepancy Δ_D between source and target. EGI reports promising transfer on synthetic and real-world data.

Adversarial training augments inputs with small, worst-case perturbations and optimizes stability to them (e.g., FGSM/PGD) (Goodfellow et al., 2015; Madry et al., 2018), encouraging invariances that often improve robustness and transfer. On graphs, analogous methods perturb structure by adding/removing a limited number of edges under a budget, which has been used to probe and enhance robustness (Zügner et al., 2018). In addition, libraries like *DeepRobust* offer an ensemble of such attacks and defenses for systematic evaluation (Li et al., 2020).

Our starting point was: if adversarial training enforces invariances in vision, can similar edge-level perturbations encourage EGI to ignore spurious graph-specific cues and generalize better? We implemented a practical, discrete *score-and-flip* procedure that perturbs edges to increase the EGI pretraining loss before fine-tuning.

We found that adversarial pretraining was inconsistent across subsets, showing gains for some and loss of performance for others. In addition to this inconsistency, we noticed a very noisy behavior across runs under the EGI framework which pushed us to understand what signal actually drives performance on these airport benchmarks. Three empirical facts emerged:

1. Degree and PageRank are almost linearly encoded in EGI embeddings; a linear probe recovers degree with $R^2 \approx 0.97$–0.99.

2. Residualizing (projecting out) degree/PageRank from embeddings causes large accuracy drops.

3. The structural distance Δ_D is not reliably predictive: Despite having worse Δ_D scores, random sources sometimes outperform the source identified as best by the original article.

Contributions. (i) Across PGD variants, adversarial pretraining consistently improved one subset and consistently hurt another, indicating context-dependent trade-offs rather than uniform benefits. (ii) Diagnostics revealing that degree/centrality constitute a major share of what EGI learns on the airport benchmarks. (iii) Indication that Δ_D alone is an unreliable pre-judge of transfer in this setting and on this benchmark.

2 Related Work

EGI and role identification. EGI models graphs as distributions over k-hop ego-structures and learns node embeddings by maximizing agreement between local encodings and a graph-level summary. The original work further proposes a structural discrepancy, denoted Δ_D, as a pre-judge for transfer—smaller Δ_D is expected to imply better source→target performance on tasks such as airport role prediction. Formally, Δ_D measures the mismatch between source and target by comparing their k-hop ego-graphs via local Laplacians. Intuitively, it is a graph-structural distance: when ego-structures align across domains Δ_D is small, and the EGI transfer gap is predicted to be small as well. Our study revisits this claim in a controlled setting that uses degree-level inputs and focuses on what signal actually drives performance(egi, 2021).

Adversarial training on graphs. In images, adversarial training generates small, worst-case perturbations (e.g., FGSM/PGD) and trains the model to keep its predictions stable—intuitively, it nudges the network to rely on *robust* features that persist under tiny pixel-level changes rather than brittle ones. Such perturbations apparently improve transferability performance as per (Salman et al., 2020). On graphs, analogous attacks perturb structure by adding/removing a few edges (or nodes) under a budget; scalable "score-and-flip" heuristics approximate the inner maximization and have been used to probe robustness and stabilize representations (Zügner et al., 2018; Li et al., 2020). We adopt a discrete score-and-flip routine during EGI pretraining and ask whether encouraging invariance to small edge edits improves cross-graph transfer in practice.

3 Background

3.1 EGI objective

Ego-Graph Infomax (EGI) treats a graph as a distribution over k-hop ego-graphs and learns node embeddings that agree with their local structure via a Jensen–Shannon MI surrogate. Let g_i be the k-hop ego-graph around node i, $z_i = f_\theta(g_i)$ its embedding, and $\mathcal{D}(g, z)$ a discriminator. The training loss is

$$\mathcal{L}_{\text{EGI}} = \frac{1}{N}\sum_{i=1}^{N}\Big[sp\big(\mathcal{D}(g_i, z_i')\big) + sp\big(-\mathcal{D}(g_i, z_i)\big)\Big] \tag{1}$$

where $sp(x) = \log(1 + e^x)$ and z_i' is a negative (mismatched) embedding drawn from another ego-graph. Intuitively, the discriminator is trained to score matching (g_i, z_i) higher than mismatched pairs.

3.2 Δ_D Guarantees

The transfer gap is not only presented as a formal bound but is also given a practical interpretation in the original EGI paper. The authors emphasize two use cases for Δ_D:

- **Point-wise pre-judge.** Given a single source graph G_a and a target graph G_b with few or no labels, one can compute $\Delta_D(G_a, G_b)$ before training to estimate whether transfer is likely to succeed. In their words: "The EGI gap $\Delta_D(G_a, G_b)$ can then be computed between G_a and G_b to prejudge whether such transfer learning would be successful before any actual GNN training (i.e., yes if $\Delta_D(G_a, G_b)$ is empirically much smaller than 1.0; no otherwise)."
- **Pair-wise pre-selection.** When multiple candidate sources $\{G_a^1, G_a^2, \dots\}$ are available, Δ_D can be computed for each pair (G_a^i, G_b) and used to select the source with the lowest Δ_D as the most promising for transfer.

Thus, Δ_D is positioned as a heuristic predictor of transfer success: lower values should imply smaller gaps and therefore better transfer. In the following sections, however, we will challenge this assumption and present cases—focusing on the airport role identification benchmarks—where smaller Δ_D does not correspond to better transfer accuracy.

3.3 Adversarial training on graphs (edge flips)

Adversarial training "finds" model weights that remain stable under worst-case, small perturbations. On graphs, a natural perturbation is to flip a few edges. Let A denote the adjacency and define the k-flip neighborhood

$$\mathcal{N}_k(A) = \{\, A \oplus S :\ S \in \{0,1\}^{n\times n},\ \|S\|_0 \le k \,\}$$

with edits being symmetric and performing at most k bit flips. A projected-gradient–style objective for pretraining is

$$\max_{A' \in \mathcal{N}_k(A)} \mathcal{L}_{\text{EGI}}(\theta; A', X),$$

after which we update θ on the adversarially edited A'.

Table 1: $\Delta_D(\text{source} \to \text{target})$ defined in EGI with EU and a random graph as sources

Source	USA	Brazil
EU	0.869	0.849
Random_{500} (BA)	0.99	0.99

We consider worst-case, small structural perturbations by flipping a few edges. Thus the maximization is approximated with a lightweight score-and-flip heuristic that ranks candidate deletions and additions and applies top-scoring edits.

4 Methodology

4.1 Score-and-Flip Adversarial Pretraining

We aim to encourage invariance to small structural edits by adversarially perturbing the source graph before standard EGI pretraining. Let A be the undirected adjacency and $\mathcal{N}_k(A)$ the neighborhood of graphs reachable via up to k edge *flips* (additions or deletions). At a high level, we approximate a discrete PGD step to find:

$$A' = argmax_{A' \in \mathcal{N}_k(A)} \mathcal{L}_{\text{EGI}}(\theta; A', X)$$

$$\Rightarrow \quad \text{train EGI on } A' \text{ with the usual recipe.}$$

Direct optimization is intractable, so we adopt a scalable greedy *score-and-flip* routine:

- Construct a small candidate set of deletions (existing edges) and additions (2-hop non-edges)
- *Score* each candidate by the marginal change in the EGI loss on a single sampled ego-batch
- *Apply* the top improving flips under soft degree caps to avoid pathological rewiring
- Iterate for a few steps with early stopping if no candidate helps

We add only *2-hop non-edges*—pairs (u, v) with $A[u, v] = 0$ but $(A^2)[u, v] > 0$ (they share a neighbor). One reason is efficiency: restricting to distance-2 candidates prunes the $O(n^2)$ non-edge space to a tractable, high-yield subset. Beyond efficiency, closing a length-2 path (triadic closure) is a realistic local edit that strongly perturbs k-hop ego-structures (the airport-role evaluation uses $k = 2$, so these flips directly affect what the discriminator observes) while avoiding long-range shortcuts and keeping global geometry stable under degree caps.

See Algorithm 1 for more details

Settings. We evaluate two PGD regimes:

- Mild: a conservative budget that perturbs local ego-structure lightly
- Aggressive: a larger budget that applies more flips and explores a larger candidate pool.

Both use the same scoring mechanism and constraints; the latter simply increases step count and candidate coverage.

4.2 Source-Variation

Since EGI (with & without PGD) yielded noisy, confusing positive and negative changes in transfer performance, we broadened our analysis to additional synthetic (relatively trivial) graphs aiming to examine the real sensitivity to the choice of source graphs, and the extent to which the structural discrepancy Δ_D reflects transferability.

4.3 Degree Diagnostics in EGI learning

Furthermore, in order to understand why PGD edits did not provide consistent improvement, we designed a small set of probes that isolate the information encoded in Z and consumed by the downstream head.

Linear degree probe. We test whether node degree is (nearly) linearly recoverable from embeddings $Z \in \mathbb{R}^{n \times h}$. We fit

$$\hat{d} \;=\; Zw + b, \qquad d := \log(1 + \deg),$$

In addition, we define a *degree–quartile prediction* task: we partition nodes into four bins by their degree, assign each node a quartile label $q \in \{1, 2, 3, 4\}$, and train a linear classifier on Z to predict q. High R^2 and quartile accuracy indicate that centrality is encoded in Z in an almost linear form and thus easily exploited by simple classifiers.

Label–degree entanglement. We measure the monotone coupling between labels and degree on each target via Spearman's ρ:

$$\rho_S \;=\; \text{Spearman}\big(y_{\text{test}},\; \deg_{\text{test}}\big).$$

Large $|\rho_S|$ explains why local edge edits that mostly preserve the majority of degree profiles may have limited downstream effect.

Table 2: Accuracy drop after residualizing centrality covariates (mean±std; paired across 100 runs).

Target	Drop (mean±std)	95% CI	Cohen's d
Europe	0.1950 ± 0.0658	$[0.1819, 0.2081]$	2.964
USA	0.1358 ± 0.0325	$[0.1294, 0.1423]$	4.181
Brazil	0.3978 ± 0.1142	$[0.3751, 0.4204]$	3.483

Residualization (ablating centrality). To quantify how much centrality drives accuracy, we remove the linear span of simple covariates. Let $C = [\mathbf{1},\ \log(1+d),\ d,\ d^{1/2},\ d^2,\ \mathrm{PR}]$ and let $\tilde{C}$ denote C z-scored column-wise. We form the orthogonal projector onto $\mathrm{span}(\tilde{C})$ via the Moore–Penrose pseudoinverse,

$$P = \tilde{C}\big(\tilde{C}^\top \tilde{C}\big)^{+}\tilde{C}^\top, \qquad Z_\perp = (I - P)\, Z.$$

We then retrain the downstream head on $Z_\perp$ and report $\Delta_{\mathrm{res}} = \mathrm{Acc}(Z) - \mathrm{Acc}(Z_\perp)$.

Degree-only baselines. We train the downstream head directly on input features X (no encoder - only degrees) to obtain Acc(X). We also compute the majority-per-degree-bin baseline: for each degree bucket b, predict the majority training label seen in that bucket and evaluate on the test set. If Acc(X) or the majority baseline approaches Acc(Z), the encoder is largely repackaging degree information rather than extracting richer transferable structure.

5 Data and Experimental Setup

5.1 Datasets & Features

We use the airport–role dataset from the EGI paper (regional subsets: EUROPE, USA, BRAZIL); these are undirected graphs with categorical role labels per airport (4 labels). Our scope follows the original EGI study, which centers on this single real-world dataset (other datasets discussed there are primarily synthetic). Although another real-world dataset has been noted, the public EGI codebase is not readily compatible with external datasets, and adapting it would require significant refactoring beyond our scope.

Features in the provided dataset are one-hot degree buckets.

5.2 Training protocol

We keep all hyperparameters and experimental methods identical to the original EGI setup. The encoder is pretrained on the *source* graph and then *frozen*. For each source→target evaluation, we fine-tune only a shallow MLP classifier on the target training split (no encoder updates), as in the EGI protocol. We repeat every experiment for 100 runs and report mean±std together with paired test statistics if applicable.

5.3 Adversarial variants

Below are the hyperparameters we use for the discrete score-and-flip routine (applied before EGI pretraining).

- **PGD (mild).** Small, localized perturbations: *steps* $= 1$, *add* $= 10$, *del* $= 10$, *candidate subset* $= 100$, *degree cap* $= 4$.
- **PGD (aggressive).** Stress-test with broader coverage: *steps* $= 3$, *add* $= 16$, *del* $= 16$, *candidate subset* $= 300$, *degree cap* $= 10$.

Hyperparameter meanings.

1. steps = number of adversarial iterations run before training
2. add/del = maximum edge additions/deletions applied per step among positively scored candidates
3. candidate subset = number of edges/non-edges randomly sampled for scoring per step
4. degree cap = a per-node limit on how much degree may change across applied flips.

5.4 Source-variation experiments

Beyond the original airport sources, we include two synthetic sources to probe sensitivity to the choice of source graph and to assess whether the structural discrepancy Δ_D reflects transferability.

Synthetic sources.

- *Random graph source:* an undirected scale-free graph generated via the Barabási–Albert model with 500 nodes (Barabási and Albert, 1999). *Degree-histogram–preserving source:* a simple undirected graph constructed with the Havel–Hakimi (Hakimi) procedure to match the degree histogram of EUROPE.

Table 3: Transfer accuracies (mean ± std) averaged over 100 runs.

Method	EU → USA	EU → Brazil
Baseline (EGI)	0.6139 ± 0.0233	$\mathbf{0.7333 \pm 0.0600}$
PGD (mild)	0.6311 ± 0.0257	0.6889 ± 0.0927
PGD (aggressive)	$\mathbf{0.6391 \pm 0.0196}$	0.6778 ± 0.0654

Table 1 presents the Δ_D between pairs of graphs showing a comparison between the random graph and Europe as sources. It was calculated using the code provided by the EGI authors.

According to the EGI interpretation, the smaller the Δ_D between a source and a target, the better the expected transfer. Refer to 3.2 for more details.

6 Results

We first compare PGD-perturbed pretraining to the baseline EGI, then unpack *why* the changes are small via degree probes, residualization, and simple baselines. Finally we revisit the Δ_D heuristic using source-variation tests.

6.1 PGD-EGI vs. EGI

We compare standard EGI pretraining to our score-and-flip PGD variants under the same encoder setup and evaluation protocol. Table 3 reports mean±std accuracies over 100 runs for EU→USA and EU→Brazil transfers.

We observe that both PGD versions provide non-negligible gain on EU→USA but degrades transfer to Brazil and increases variance. Another interesting observation is that with more aggressive edits, EU → USA improves while transferability to Brazil goes down. Given the mixed effect, we focus next on diagnosing why transfer is both inconsistent and noisy.

Table 4: Degree-probe statistics (mean over 100 runs).

Dataset	$\lvert\rho\rvert$(label,deg)	R^2	Quartile-acc
Europe	0.79	0.99	0.999
USA	0.77	0.97	0.971
Brazil	0.88	0.99	1.000

6.2 Degree is linearly encoded

We probe whether centrality is explicitly stored in the learned embeddings. For each run, we (i) measure label–degree coupling on the *target test* split (Spearman), and (ii) regress $\log(1 + \deg)$ from Z and classify degree quartiles with a linear head. Table 4 contains the Spearman(label,deg), $R^2(Z \to \log(1 + \deg))$, and Quartile accuracy averaged over a 100 runs.

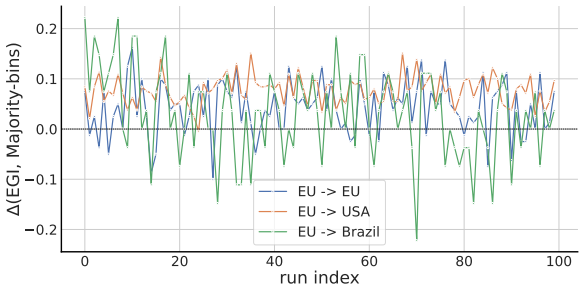

Figure 1: Per-run Δ between EGI and a majority-per-degree-bin rule.

Taken together, large $|\rho|$, $R^2 \approx 1$, and near-perfect quartile accuracy indicate that Z makes degree linearly accessible. This raises a shortcut risk: a simple MLP head can rely on degree (and closely related centrality) rather than higher-order structure, which helps explain why degree-preserving edge flips yield only modest transfer changes.

6.3 Residualizing degree/PageRank hurts

The previous experiment showed that degree is almost linearly encoded in Z. We now ask whether the degree/centrality signal is causal for transfer accuracy by projecting Z onto the orthogonal complement of simple centrality covariates (degree transforms and PageRank), then re-training only the target MLP head. Table 2 reports the paired accuracy drop across 100 runs.

As seen in Table 2, drops are universal, large, and statistically strong. Brazil is most degree-dependent, consistent with its higher label–degree coupling. Together with the probe results, this supports the view that much of EGI's transfer signal on airports is mediated by simple centrality.

6.4 Simple degree baselines are strong

Having shown that degree is both linearly encoded in Z and causally linked to accuracy, we now directly compare a trivial degree-aware classifier—*majority-per-degree-bin* (predict the training-set majority label within each degree bin)—against EGI. Figure 1 visualizes the per-run gap $\Delta = \mathrm{Acc}(Z) - \mathrm{Acc}(\mathrm{majority})$, and Table 5 summarizes the paired statistics over 100 runs.

As Figure 1 and Table 5 show, gains over a trivial degree baseline are limited for Brazil and Europe but sizable and more consistent for USA. Importantly, the majority-per-degree-bin rule is far more

Table 5: EGI vs. majority-per-degree-bin (EU source; $n = 100$). $\Delta = \mathrm{Acc}(Z) - \mathrm{Acc}(\text{majority})$.

Target	mean Δ	95% CI	$\% \Delta{>}0$	Cohen's d
Brazil	$+0.031$	$[0.013, 0.048]$	58%	0.35
Europe	$+0.043$	$[0.032, 0.053]$	75%	0.80
USA	$+0.077$	$[0.071, 0.083]$	99%	2.63

efficient than EGI: it needs no encoder pretraining or contrastive batches—just a single pass to tally bucket-wise majorities—so its small gap on Brazil/Europe comes at a fraction of EGI's compute. By contrast, on USA there is clear headroom beyond the degree rule, and—interestingly—our PGD-EGI variant outperforms EGI consistently on this subset, suggesting additional transferable cues may be exposed by local edge edits; we leave a deeper analysis of this correlation to future work.

6.5 Source variation and Δ_D

Here we revisit the Δ_D pre-selection idea by swapping the source (i.e., the original EUROPE), with a random graph and a degree-preserving graph and report the results in Table 6.

We can see that degree-preserving source beats EUROPE on BRAZIL, while a random graph is best for EU→USA. Thus, the most "similar" source by global structure is not consistently best.

7 Conclusions

We set out to test whether adversarial edge edits can enhance cross-graph transfer for EGI on airport-role benchmarks. Across 100-run evaluations, PGD-perturbed pretraining produced mixed effects: a non-negligble gain on EU→USA, degradation with higher variance on EU→Brazil. Simple diagnostic tests reveal a possible explanation: degree/centrality is (almost) linearly encoded in the embeddings and is strongly entangled with labels. Residualizing degree/PageRank causes large, universal drops—especially on BRAZIL—indicating that much of the transferable signal is mediated by simple centrality. Source-variation tests further show that the "closest" source by global structure need not be best, cautioning against relying on Δ_D alone for pre-selection.

8 Limitations

Adversarial scoring under compute constraints. Our score-and-flip routine scores candidates on a single ego-batch per step. This introduces coverage gaps and noisy Δ estimates that can miss globally meaningful flips. The choice is pragmatic: the public EGI code depends on legacy graph packages that were incompatible with our university GPUs; refactoring them for modern CUDA stacks proved non-trivial despite attempts. As a result, most experiments ran on CPU with strict wall-clock limits, constraining candidate set sizes and PGD step budgets (timings in the Appendix).

Dataset and code coupling. Findings are reported on the airport-role benchmarks because the released EGI repository is tightly coupled to these graphs. Incorporating external real-world networks would require substantial engineering; we attempted to adapt the code and sought guidance on the repository but received no response. Consequently, some observations may reflect properties of this specific benchmark (e.g., its degree–label coupling and preprocessing) rather than EGI in general. Conclusions should be read in this context; broader validation on additional datasets is important future work.

Table 6: EGI Transfer accuracies for alternative sources.

Source → Target	EU	USA	Brazil
Europe	$\mathbf{0.5755 \pm 0.0496}$	0.6151 ± 0.0302	0.7204 ± 0.0769
Random	0.5665 ± 0.0467	$\mathbf{0.6289 \pm 0.0295}$	0.6596 ± 0.0829
Degree-Preserving Eu.	0.5518 ± 0.0524	0.6053 ± 0.0307	$\mathbf{0.7622 \pm 0.0675}$

References

2021. Ego-graph information maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Seifollah Louis Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph ii. uniqueness. *Journal of the Society for Industrial and Applied Mathematics*, 11(1):135–147.

Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. 2020. Deeprobust: A pytorch library for adversarial attacks and defenses. Includes an ensemble of attacks/defenses for graphs and images.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. 2020. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2847–2856. ACM.

A PGD-style score-and-flip pseudocode

Algorithm 1 Score-and-Flip (one adversarial pretraining phase)

Require: adjacency A, features X, step budget T, flip limits $(k_{\text{add}}, k_{\text{del}})$, degree cap $d_{\max}$

for $t = 1, 2, \ldots, T$ **do**
  Sample a small ego-batch $\mathcal{B}$ for loss estimation
  Build candidate sets: additions $\mathcal{C}_{\text{add}}$ (2-hop non-edges), deletions $\mathcal{C}_{\text{del}}$ (existing edges)
  for all $(u, v) \in \mathcal{C}_{\text{add}} \cup \mathcal{C}_{\text{del}}$ **do**
    Estimate marginal gain $\Delta(u, v) \leftarrow \mathcal{L}_{\text{EGI}}(A \oplus \text{flip}(u, v); X \mid \mathcal{B}) - \mathcal{L}_{\text{EGI}}(A; X \mid \mathcal{B})$
  end for
  Select up to k_{add} positive-gain additions and k_{del} positive-gain deletions
  Apply flips that satisfy symmetry and soft degree constraints ($\deg \leq d_{\max}$)
  if no flips applied **then break**
  end if
end for
Train EGI on the perturbed graph A' with the standard (clean) procedure