POSITIVE MINING FROM LLM SEEDS: A SEMI-SUPERVISED GRAPH BASED APPROACH TO TRAIN RARE EVENT CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting rare events, from emerging hate speech to novel fraud patterns, presents a fundamental cold-start challenge: without labeled examples, we cannot train classifiers, and manually searching vast unlabeled corpora for rare instances is prohibitively expensive. This paper introduces SYNAPSE-G (Synthetic Augmentation for Positive Sampling via Expansion on Graphs), a framework that bridges Large Language Models and graph-based learning to efficiently bootstrap rare event detection from scratch. Rather than using synthetic data for direct model training, SYNAPSE-G employs LLM-generated examples as intelligent "seeds" to efficiently probe large unlabeled datasets. These seeds initialize a semi-supervised label propagation process over a similarity graph, identifying real candidate instances for oracle verification. We provide a theoretical analysis connecting the quality of synthetic seeds, specifically their validity (accuracy) and diversity (coverage), to the precision and recall of discovered positives, revealing a nuanced trade-off between these properties. Through systematic evaluation on imbalanced SST2 and Measuring Hate Speech datasets, we demonstrate that SYNAPSE-G discovers 28.6% of rare positives while querying only 2.4% of data, substantially outperforming standard active learning baselines. Our work establishes design principles for combining synthetic data generation with graph-based discovery in extreme class imbalance scenarios.

1 INTRODUCTION

The detection of rare events represents a critical challenge across diverse domains: identifying fraudulent transactions in financial systems (Bauder & Khoshgoftaar, 2020; Carreño et al., 2020), diagnosing rare diseases in healthcare (Shyalika et al., 2024), and combating emerging forms of online abuse tied to current events. This challenge intensifies with the rapid evolution of internet trends such as novel misinformation tactics (Suarez-Lledo & Alvarez-Galvez, 2021), new slang used for harassment (Mathew et al., 2019; Siegel, 2020), or emerging fraud schemes (Herland et al., 2019), where labeled training data is inherently scarce due to the phenomenon's novelty (Shyalika et al., 2024).

This scarcity creates a vicious cycle: without labeled examples, we cannot train effective classifiers, yet without classifiers, we cannot efficiently locate rare examples within massive unlabeled corpora. For instance, to build a detector for a new harassment term, one would typically need to manually review millions of social media posts to find a handful of positive instances. This cold-start problem becomes especially acute when timely detection is critical, such as responding to hate speech targeting specific groups during unfolding events.

Recent advances in Large Language Models (LLMs) offer a potential solution: these models can generate synthetic examples of rare events based on textual descriptions alone Ding et al. (2024). However, prior work has primarily focused on using synthetic data as direct training material for downstream classifiers Chen et al. (2023); Ye et al. (2022). This approach faces two fundamental limitations: (1) synthetic data exhibits redundancy and distribution mismatch with real-world instances Chen et al. (2024), and (2) models trained solely on synthetic data often fail to generalize to authentic examples Seddik et al..

A New Paradigm: Synthetic Data as Discovery Seeds. We propose a fundamentally different use of synthetic data. Rather than training models directly on LLM-generated examples, we use them as intelligent probes to efficiently discover real positive instances within large unlabeled datasets. Our framework, SYNAPSE-G (Synthetic Augmentation for Positive Sampling via Expansion on Graphs), treats synthetic examples as "seeds" that bootstrap a graph-based semi-supervised learning process.

The key insight is that while synthetic data may not perfectly replicate real-world distributions, it provides semantic guidance about what the rare event looks like in embedding space. By constructing a similarity graph over synthetic seeds and unlabeled real data, we can propagate labels through the graph structure to identify strong candidates for oracle verification. This approach sidesteps the challenges of direct synthetic training while leveraging LLMs' ability to generate diverse examples from minimal descriptions.

1.1 Overview of Synapse-G

Our framework operates in four stages:

1. **Synthetic Seed Generation:** An LLM generates a pool of synthetic rare event examples. We select a small, diverse subset using either random sampling or Adaptive Coverage Sampling (ACS) (Tavakkol et al., 2025), which maximizes coverage of the semantic space.
2. **Graph-Based Candidate Discovery:** We embed synthetic seeds and unlabeled real data into a shared semantic space, constructing a similarity graph. Label propagation from seeds identifies real instances with high structural similarity to synthetic examples.
3. **Oracle Labeling:** Top candidates are presented to an oracle (human annotator or high-quality LLM) for verification, rapidly building a labeled set of authentic positives.
4. **Model Training:** A classifier is trained or fine-tuned on the verified real examples for deployment.

We explore two graph propagation strategies: Iterative Bipartite Graph (IBG), which performs local expansion from known positives, and Graph-Based Label Expansion (GBLE), which leverages global graph structure through standard label propagation (Zhu & Ghahramani, 2002).

1.2 Theoretical Insights

A central contribution of this work is a formal analysis of how synthetic seed quality impacts discovery performance. We introduce two quality dimensions: (1) validity (p) or the proportion of synthetic seeds that truly represent positive instances, and (2) diversity (h) or the coverage of the seed set over the data distribution (vertex expansion ratio).

Through analysis on a simplified graph model, we prove that while recall monotonically increases with both validity and diversity, precision exhibits a non-trivial interaction between these properties (Proposition 4.1). Specifically, there exists a validity threshold above which increasing diversity decreases precision, and below which diversity increases precision. This reveals that optimal seed selection must balance these competing objectives based on the expected quality of LLM-generated data.

This theoretical framework provides actionable guidance: when LLM-generated seeds are highly reliable, focus on finding overlapping neighborhoods (lower diversity, higher precision via multiple confirmations). When seed reliability is uncertain, cast a wider net (higher diversity, accepting lower precision per query but achieving better coverage).

1.3 Empirical Validation and Key Findings

We validate SYNAPSE-G on two tasks with different characteristics.

Controlled Evaluation (SST2). On an artificially imbalanced sentiment dataset (10% positive), we systematically compare design choices. Most crucially, we observe that ACS seed selection consistently outperforms random selection across all metrics. Moreover, our GBLE propagation method significantly outperforms the more naive IBG in iterative discovery scenarios. As a result,

with optimal configuration, Synapse-G achieves 40.8% retrieval of the rare data while only querying 5% of the data for labels.

Realistic Rare Event (Measuring Hate Speech). The second task is positioned for detecting hate speech targeting transgender individuals (6.5% base rate, only 19 synthetic seeds available in the MHS dataset of Sachdeva et al. (2022)). For this setting, SYNAPSE-G discovers: 28.6% of true positives with only 2.4% of labels queried, 40.8% of true positives with only 5.0% of labels queried, and the method maintains higher recall than a natural logistic regression baseline *even when this baseline has access to known negative examples and large inference budgets*.

These results demonstrate that graph-based discovery with synthetic seeds provides a practical, scalable approach to cold-start rare event detection, substantially reducing labeling costs compared to random sampling or standard active learning approaches.

1.4 Our Contributions

Overall, this work makes the following contributions. We first introduce the paradigm of using synthetic data as discovery seeds rather than direct training material, addressing the cold-start problem in rare event detection. We subsequently provide formal analysis (Proposition 4.1) connecting synthetic seed quality (validity and diversity) to discovery performance (precision and recall), revealing a nuanced trade-off between these properties. Through systematic comparison of seed selection strategies (random vs. ACS) and propagation methods (IBG vs. GBLE), we establish that ACS + GBLE provides the most effective configuration for rare event discovery. We thus proceed to demonstrate SYNAPSE-G's effectiveness on both controlled benchmarks and realistic rare event tasks, showing substantial improvements over baselines in label efficiency.

The remainder of this paper is organized as follows: Section 2 formalizes the rare event detection problem. Section 3 details SYNAPSE-G's methodology. Section 4 presents our theoretical analysis supplemented by Section 5 which reports experimental results. We defer the comprehensive review of related work in retrieval, active learning and graph-based semi-supervised learning to Appendix A due to space constraints.

2 Preliminaries & Problem Definition

We formalize rare event detection as a binary classification task under two critical constraints: (1) severe class imbalance and (2) complete absence of initial labeled data (the cold-start problem).

2.1 Setup and Notation

Let $\mathcal{X}$ denote the input space, where each observation $x \in \mathcal{X}$ has an associated binary label $y \in \{0, 1\}$. Here $y = 1$ indicates a rare positive instance (the event of interest) and $y = 0$ indicates a common negative instance. The data is drawn from an underlying distribution $\Pr(x, y)$ with severe class imbalance: $\Pr(y = 1) \ll 0.5$.

Initially, we are given a large unlabeled dataset $\mathcal{D}_U = \{x_j\}_{j=1}^N$ with N instances, where labels $\{y_j\}_{j=1}^N$ are unknown. Critically, we have zero labeled positive examples initially (the cold start constraint). We have access to an oracle $\mathcal{O}$ (human annotator or high-quality LLM) that can provide labels for queried instances. Each query has a cost, so our goal is to minimize the total number of oracle calls while maximizing discovery of rare positives. To bootstrap the discovery process, we assume access to a mechanism (typically an LLM) that can generate synthetic examples of the rare event given a textual description. Let $\mathcal{D}_S = \{\tilde{x}_i\}_{i=1}^{n_s}$ denote this synthetic dataset.

2.2 Active Learning Procedure

We operate within an iterative active learning framework over T rounds. In each iteration $t \in \{1, \ldots, T\}$ a strategy π initially selects a batch $\mathcal{B}_t \subseteq \mathcal{D}_U$ of size B (the labeling budget per iteration). Next, the oracle provides labels for the batch, yielding $\mathcal{L}_t = \{(x, y) : x \in \mathcal{B}_t\}$. The strategy updates its internal state (e.g., set of known positives, model parameters) based on $\mathcal{L}_t$ and previous labeled sets $\bigcup_{j=1}^{t-1} \mathcal{L}_j$ Lastly, we remove the newly labeled data $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{B}_t$.

After T iterations, we have labeled $\mathcal{L}^{(T)} = \bigcup_{t=1}^{T} \mathcal{L}_t$ containing $T \cdot B$ instances.

2.3 Evaluation Metrics

The quality of a selection strategy π is measured by its ability to efficiently discover rare positives. We use two complementary metrics evaluated at each iteration $k \in \{1, \ldots, T\}$. We define the *cumulative precision* as the fraction of all queried instances (up to iteration k) that are positive:

$$P^{(k)} = \frac{|\{(x, y) \in \mathcal{L}^{(k)} : y = 1\}|}{|\mathcal{L}^{(k)}|} = \frac{|\{(x, y) \in \mathcal{L}^{(k)} : y = 1\}|}{k \cdot B}$$

High precision indicates the strategy has a high "hit rate", or that most oracle queries are productive, discovering true positives rather than wasting effort on negatives. We define the *cumulative recall* as the fraction of all true positives in $\mathcal{D}_U$ that have been discovered by iteration k:

$$R^{(k)} = \frac{|\{(x, y) \in \mathcal{L}^{(k)} : y = 1\}|}{n^{+}_{total}}$$

where $n^{+}_{total} = |\{x \in \mathcal{D}_U : y = 1\}|$ is the (unknown) total number of positives in the original unlabeled set. High recall indicates the strategy has found a large fraction of the rare events. An ideal strategy maximizes both precision and recall at every iteration. Since n^{+}_{total} and B are fixed constants, maximizing $P^{(k)}$ and $R^{(k)}$ is equivalent to maximizing the cumulative count of discovered positives:

$$n^{+}_{discovered} = |\{(x, y) \in \mathcal{L}^{(k)} : y = 1\}| = P^{(k)} \cdot k \cdot B = R^{(k)} \cdot n^{+}_{total}$$

In practice, there is often a trade-off: strategies that query conservatively (high precision) may miss positives (low recall), while aggressive strategies (high recall) may waste oracle budget on negatives (low precision). Section 4 analyzes this trade-off theoretically for graph-based strategies using synthetic seeds.

2.4 The Cold-Start Challenge

The key challenge distinguishing our setting from standard active learning is the cold-start problem where we have zero real labeled positives initially ($\mathcal{L}^{(0)} = \emptyset$). Traditional query strategies require an initial model trained on labeled data. A naive solution to overcoming this barrier is random sampling until finding enough positives to train an initial model. For a rare event with prevalence $p = \Pr(y = 1)$, we need to label approximately k/p instances to find k positives. With $p = 0.01$ (1% prevalence) and needing $k = 20$ positives, this requires labeling 2,000 instances.

Instead, our proposed method uses LLM-generated synthetic positives $\mathcal{D}_S$ as semantic anchors in embedding space. These synthetic seeds bootstrap a graph-based discovery process that identifies real candidates without requiring a trained model. This enables efficient discovery from iteration 1, sidestepping the cold-start problem entirely.

3 Methodology

We now present SYNAPSE-G, a framework that leverages synthetic data and graph structure to bootstrap rare event discovery without initial real labels.

3.1 Synthetic Seed Generation

Starting from zero real labels, we use an LLM to generate synthetic positive examples that serve as a bootstrap for the discovery process. Given a textual description of the rare event (e.g., "hate speech targeting transgender individuals"), an LLM generates a pool of synthetic positive examples $\mathcal{D}_S^{pool}$. While sophisticated prompt engineering significantly impacts synthetic data quality (Liu et al., 2024), our focus is on the downstream utilization of synthetic data for discovery. We therefore use publicly available pre-generated synthetic datasets (**?**Casula et al., 2024) to isolate the effect of our selection and propagation strategies from prompt engineering choices.

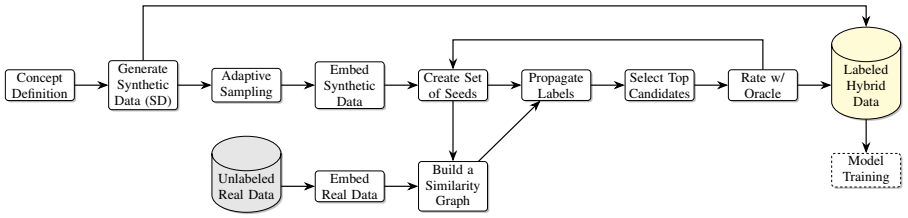


Figure 1: Overview of the SYNAPSE-G pipeline for rare event classification. The pipeline integrates synthetic data generation (top branch) with real data processing (bottom branch). LLM-generated synthetic data, after adaptive sampling and embedding, forms a set of positive seeds. Unlabeled real data is also embedded. A similarity graph connects seeds and real data, enabling label propagation to identify top candidates. An oracle rates these candidates, creating a labeled hybrid dataset for model training. The iterative nature is shown by the feedback loop from "Rate w/ Oracle" to "Create Set of Seeds".

From the pool $\mathcal{D}_S^{pool}$, we select a small, diverse subset $\mathcal{D}_S \subset \mathcal{D}_S^{pool}$ with $|\mathcal{D}_S| = n_s$ to serve as initial seeds. We compare two selection strategies. The naive approach is random sampling, wherein we uniformly sample n_s examples from $\mathcal{D}_S^{pool}$. This serves as a baseline, requiring no additional computation. We additionally leverage ACS (Tavakkol et al., 2025) where we select n_s examples that maximize coverage of the semantic space (proportion of the pool "represented" by the selected subset). ACS frames this as a maximum coverage problem on a similarity graph, using binary search to find the optimal similarity threshold and greedy approximation for selection. This ensures diversity since seeds span different semantic regions rather than clustering around common patterns. Our theoretical analysis (Section 4) reveals that seed diversity directly impacts recall, while interacting non-trivially with validity to affect precision. ACS explicitly maximizes diversity, which we hypothesize will improve discovery performance when synthetic seeds are reasonably valid. Figure 1 illustrates the complete pipeline.

3.2 Embedding and Similarity Graph Construction

All data points, both synthetic seeds $\mathcal{D}_S$ and unlabeled real data $\mathcal{D}_U$, are mapped into a shared semantic space using a pre-trained embedding model. In our experiments, we use Gecko embeddings (Lee et al., 2024), though SYNAPSE-G is compatible with any embedding function.

We construct an undirected similarity graph $G = (V, E)$ where vertices represent all synthetic seeds and unlabeled real data ($V = \mathcal{D}_S \cup \mathcal{D}_U$) and edges connect pairs (x_i, x_j) with $\text{sim}(x_i, x_j) \geq \tau$ for threshold τ. The threshold τ controls graph density. Higher τ creates sparser graphs (only very similar instances connect), while lower τ creates denser graphs (more connections, including weaker similarities). Our graph-based strategies explore this trade-off.

3.3 Graph-Based Candidate Discovery

Given the similarity graph with synthetic seeds as initial positive labels, we identify real instances likely to be positive through graph-based label propagation. We propose and compare two complementary strategies: one local and iterative (IBG), one global and batch-oriented (GBLE).

Iterative Bipartite Graph (IBG). IBG performs local expansion from known positives. Specifically, it constructs a bipartite graph between the current set of confirmed positives V_P (initially $\mathcal{D}_S$) and remaining unlabeled data V_U, then selects the unlabeled instances most similar to positives. In each iteration t, we first construct the bipartite graph $G_B = (V_P, V_U, E_B)$ where edges connect (v_i, v_j) with $v_i \in V_P$, $v_j \in V_U$, and $\text{sim}(v_i, v_j) > \tau$. For each $v_i \in V_P$, we retain only its top $d_{\max}$ neighbors in V_U (by similarity) to prevent high-degree samples from dominating selection and encourage diversity. The procedure then selects all unlabeled nodes connected to at least one positive as the candidate set, and obtains labels from the oracle. The set of confirmed positives is then added to the set of correctly labeled samples. We defer the full pseudocode to Appendix B (Algorithm 1).

Note that IBG is conservative in that it only explores the immediate neighborhood of known positives. This yields high precision (selected instances are very similar to confirmed positives) but may

limit recall if the positive distribution is fragmented or if synthetic seeds do not cover all positive regions.

Graph-Based Label Expansion (GBLE). GBLE performs global label propagation over the entire similarity graph. Rather than iteratively expanding from positives, it propagates positive scores through the graph structure, leveraging both positive and negative labels to refine estimates.

The procedure first initializes labels where synthetic seeds are assumed positive and unlabeled instances have no assignment. We then construct an adjacency matrix and diffuse label information through the graph via a random walk transition for T_{prop} iterations (node labels are the weighted average of their neighbors). We then rank unlabeled nodes by their final positive score nd select the top k as candidate batch $\mathcal{B}_t$. Unlike IBG (fixed batch size), GBLE adjusts k based on previous precision: $k = \lceil k_0/p_{prev} \rceil$ where k_0 is target number of positives to find and p_{prev} is precision from iteration $t-1$. If precision is high (finding many positives), query fewer instances. If precision is low (mostly negatives), cast a wider net. After the oracle call for labels $\mathcal{B}_t$, we reset the based set of assumed positives and repeat. Algorithm 2 formalizes the label propagation subroutine.

GBLE is exploratory and leverages the global graph structure to identify candidates that may be structurally similar to positives even if not directly connected. The dynamic batch sizing further provides adaptive exploration. In Section 5 we test the hypothesis that GBLE will outperform IBG in iterative discovery since GBLE uses labeled negative to refine boundaries and further leverages a global, adaptive, structural approach.

4 Theoretical Analysis

Having described SYNAPSE-G's methodology, we now provide theoretical foundations for understanding when and why synthetic seeds enable effective rare event discovery. Specifically, we analyze how validity (accuracy) and diversity (coverage) jointly determine the precision and recall of the discovery process. While intuition suggests that more diverse seeds always improve performance, our analysis reveals a non-monotonic relationship: diversity can help or hurt precision depending on seed validity.

4.1 Model and Assumptions

To make the analysis tractable, we study a simplified single-iteration version of our method on an idealized graph structure. While real data does not perfectly match these assumptions, the model provides qualitative insights that guide empirical design choices. We represent data as an undirected, d-regular graph $G = (V, E)$ where each node $v \in V$ represents a data point with true label $y(v) \in \{0, 1\}$ ($y(v) = 1$ for positive instances). Each node has exactly degree d, and edges indicates semantic similarity in embedding space

The high-level algorithm proceeds as follows: given synthetic seed set $S \subset V$, we query all seeds S and obtain their true labels, partitioning into $S_+ = \{v \in S : y(v) = 1\}$ (true positives) and $S_- = \{v \in S : y(v) = 0\}$ (false positives) We then query all immediate neighbors of true positive seeds: $N(S_+) = \bigcup_{v \in S_+} N(v)$ where $N(v) = \{u : (u, v) \in E\} \cup \{v\}$ (neighbors plus self) to obtain the queried set $Q = S \cup N(S_+)$. This mirrors the first iteration of IBG (Algorithm 1) with threshold τ chosen such that only direct neighbors are included. We proceed to make simplifying assumptions about seed quality and graph structure.

Assumption 1 (Seed Diversity)**.** *S is well-dispersed in the graph: S forms an independent set (no edges between seed nodes) and no node in V is adjacent to more than two seed nodes*

This formalizes the notion that seeds are spread out rather than clustered. The first condition ensures seeds are distinct (no redundancy), while the second prevents any single node from being overwhelmed by multiple seed influences. Observing that the seed set can always be partitioned into true positives, $S_+ = \{v \in S : y(v) = 1\}$ and false positives $S_- = \{v \in S : y(v) = 0\}$, , we define validity as $p = \frac{|S_+|}{|S|}$ the proportion of synthetic seeds that are truly positive. Validity captures the accuracy of synthetic data generation. High p means the LLM produces mostly correct positive examples; low p means many false positives contaminate the seeds.

We further quantify seed diversity via the vertex expansion ratio, $h(S) = \frac{|N(S)|}{|S|} \geq 1$ This measures how many distinct nodes are reachable from S (coverage). Higher $h(S)$ indicates seeds reach more of the graph. For a d-regular graph, $1 \leq h(S) \leq d + 1$.

Assumption 2 (Graph Homophily)**.** *The probability that an unlabeled node u is positive depends on how many positive seeds it connects to. Specifically, if u is adjacent to exactly n positive seeds in S_+ (for $n = 1$ or $n = 2$), then:*

$$\Pr[y(u) = 1 \mid u \text{ adjacent to } n \text{ positives in } S_+] = q_n$$

where $0 < q_1 < q_2 < 2q_1$.

Intuitively, q_1 is the probability that a neighbor of one positive seed is itself positive (homophily strength). q_2 is the probability for neighbors of two positive seeds (confirmation from multiple sources). The constraint $q_1 < q_2 < 2q_1$ ensures that double confirmation helps, but is not simply additive which is consistent with probabilistic independence.[1] Under these assumptions, we characterize the expected precision and recall of the discovery process.

Proposition 4.1. *Let $Q = S \cup N(S_+)$ be the set of queried nodes and P the number of true positives in Q. Then the expected precision and recall, conditioned on the seed set S, are:*

$$\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h(S_+)}$$

and ***Recall:***

$$\mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\left((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\right)$$

respectively.

We defer the proof to Appendix C for a complete derivation.

4.2 Interpretation and Practical Insights

Proposition 4.1 reveals several non-obvious insights about synthetic seed quality and discovery performance. First, recall is monotonically increasing in both validity (p) and diversity ($h(S_+)$). This aligns with expectation—better seeds (higher p) and more coverage (higher $h(S_+)$) allow discovering more of the total positive population. Since $(2q_1 - q_2) > 0$ (from $q_2 < 2q_1$), the coefficient on $h(S_+)$ is positive, so diversity always helps recall. When the goal is maximizing recall (e.g., finding as many rare instances as possible regardless of precision), use ACS to select maximally diverse seeds.

We further observe that precision's dependence on diversity ($h(S_+)$) is non-monotonic. Taking the derivative of precision with respect to $h(S_+)$:

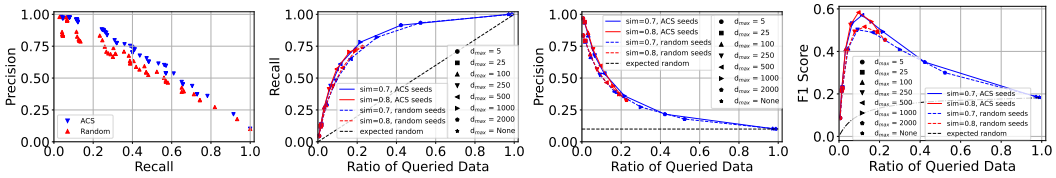
$$\frac{\partial}{\partial h(S_+)}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = -\frac{1 + q_2(d + \frac{1}{p}) - q_1(d + \frac{2}{p})}{(\frac{1-p}{p} + h(S_+))^2}$$

The sign depends on the numerator: $1 + q_2(d + \frac{1}{p}) - q_1(d + \frac{2}{p})$. This is positive (precision decreases with diversity) when:

$$p > p^* = \frac{2q_1 - q_2}{1 + (q_2 - q_1)d}$$

This defines two regimes for high and low validity. In the high validity regime ($p > p^*$), precision decreases with diversity. When synthetic seeds are highly accurate (large p), the main challenge is distinguishing true positives from structurally similar negatives. In this regime, neighbors of a single positive seed have modest positive probability q_1, whereas neighbors of multiple positive seeds

[1] If each positive seed independently induces a positive label with probability q_1, then $q_2 = 1 - (1 - q_1)^2 = 2q_1 - q_1^2$, satisfying the constraint.



(a) Precision vs. Recall (b) Recall vs. Query. (c) Precision vs. Query. (d) F1 Score vs. Query.

Figure 2: Experimental results on single-shot evaluations on the imbalanced SST2 dataset. (a) Precision vs. Recall, comparing ACS and random seed selection. (b) Recall vs. Query Ratio, showing the benefit of ACS seeds. (c) Precision vs. Query Ratio, illustrating the impact of parameters. (d) F1 Score vs. Query Ratio, showing the impact of parameters. ACS consistently outperforms random seed selection across all metrics. Figures (b), (c), and (d) demonstrate the impact of varying similarity thresholds and maximum degree constraints. Curves above the expected random performance (diagonal in (b) and horizontal/curve in (c)/(d) respectively) indicate a benefit of the graph-based approach.

(overlapping neighborhoods) have higher probability q_2. Therefore, focusing queries on regions where seed neighborhoods overlap (lower $h(S_+)$) yields higher precision through "confirmation by multiple sources"

In the low validity regime ($p < p^*$), precision increases with diversity. When synthetic seeds have low accuracy (small p), many seeds are false positives. Thus, the false positive seeds S_- contribute noise, and their neighborhoods are likely negative, while maximizing diversity (higher $h(S_+)$) spreads the true positive seeds S_+ widely, covering more distinct positive regions while minimizing redundant coverage of negative regions.

The threshold, $p^* = \frac{2q_1 - q_2}{1+(q_2-q_1)d}$ increases with q_1 (stronger single-neighbor signal) and decreases with q_2 (stronger double-neighbor signal) and d (more connections per node). From this, we lastly observe that precision monotonically increases with validity p regardless of diversity. Computing the derivative:

$$\frac{\partial}{\partial p}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = \frac{1 + d(q_2 - q_1) + (h(S_+) - 1)(2q_1 - q_2)}{(\frac{1-p}{p} + h(S_+))^2} > 0$$

which is always positive since all terms are non-negative and $h(S_+) \geq 1$. This suggests that improving synthetic data generation quality (higher p) unambiguously helps precision.

5 Experimental Results

We evaluate SYNAPSE-G through a two-phase experimental design that progressively validates our approach: (1) controlled analysis on artificially imbalanced SST2 to systematically compare design choices and test theoretical predictions from Section 4, and (2) realistic evaluation on naturally imbalanced MHS to demonstrate practical effectiveness against strong baselines.

5.1 Experimental Setup

All experiments use pre-trained Gecko embeddings (Lee et al., 2024) (768 dimensions), to map text into semantic space. Cosine similarity between embeddings defines edge weights in similarity graphs. Following the prior work, we evaluate methods using the cumulative precision $P^{(k)}$ (fraction of queried instances that are positive), cumulative recall $R^{(k)}$ (fraction of true positives discovered), F1 score, and the query ratio (fraction of unlabeled data labeled). For single-shot IBG experiments, we vary the similarity threshold for $\tau \in \{0.7, 0.8\}$ and the maximum degree $d_{\max} \in \{5, 25, 100, 250, 500, 1000, 2000, \text{None}\}$ For GBLE, we use $T_{prop} = 5$ propagation iterations and $K_0 = 100$ target positives per round.

5.2 Controlled Analysis on SST2

Dataset: The Stanford Sentiment Treebank 2 (SST2) (Socher et al., 2013) contains 67,349 movie review sentences with binary sentiment labels (positive/negative). We use a publicly available syn-

thetic SST2 dataset generated by GPT-3.5 (Ding et al., 2023) containing 5,000 synthetic reviews. To mimic extreme class imbalance, we subsample the positive class to create an imbalanced training set. From the real dataset, we keep all 29,780 negatives and subsample to 3,308 positives (10% positive rate). We select 100 positive seeds from 2,488 synthetic positives to probe for the real positives. Table 1 summarizes the dataset composition.

Experimental conditions: We evaluate the two seed selection strategies of random sampling and ACS with coverage $c = 0.5$, each combined with graph-based discovery using IBG (Algorithm 1) and GBLE (Algorithm 2).

We first analyze single-iteration performance (one round of candidate selection and labeling) to test predictions from Section 4. As baselines, we use: (1) random selection, which captures the expected performance assuming uniform random sampling, (2) random seeds + IBG, where we obtain 100 randomly selected synthetic seeds with IBG propagation, and lastly (3) ACS Seeds + IBG which uses 100 ACS-selected seeds with IBG propagation Figure 2 summarizes results across four metrics. We see that ACS consistently outperforms random seed selection. The precision-recall curve shows ACS achieves 5-10% higher precision than random seeds at matched recall levels. This validates our prediction that ACS's coverage maximization improves discovery efficiency. Moreover, we observe that graph-based methods dramatically outperform the random baseline. Both ACS and random seeds with IBG propagation discover the majority of positives with only 20% query ratio, and 30-50% precision at low-query ratios. This confirms our prediction that graph structure enables efficient discovery by exploiting semantic similarity. The dashed lines in Figure 2 show expected random performance, with all graph-based curves lie substantially above. We further see that parameter effects align with theoretical insights. Specifically, performance improves with $d_{\max}$ but with diminishing returns. Lastly, we note that F1 score reveals optimal operating points, peaking at 20-30% query ratio for most configurations (with ACS consistently achieving higher peaks, justifying its use for seed selection).

5.3 Iterative Evaluation: IBG vs. GBLE

Section 3 hypothesized that GBLE should outperform IBG in multi-iteration scenarios due to the utility of negative samples and global approaches. We test this by running both methods iteratively (up to 10 rounds) where IBG operates on a fixed batch size, $\tau = 0.8$, $d_{\max} \in \{5, 25, 100\}$. GBLE instead uses dynamic batch sizing ($K = K_0/p_{prev}$, $K_0 = 100$), $T_{prop} = 5$. These are contrasted against an oracle with perfect precision (upper bound on recall curve).

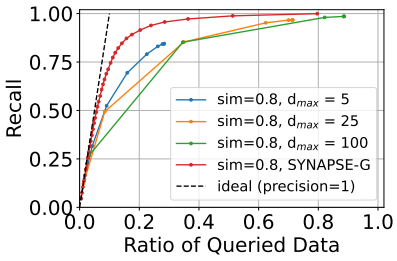


Figure 3: Recall vs. Ratio of Queried Data for *iterative* rare event detection on the imbalanced SST2 dataset. "Ideal" is perfect precision. The graph-Based Label Expansion (GBLE) significantly outperforms Iterative Bipartite Graph (IBG).

GBLE achieves 90% recall with 40% query ratio (13,000 queries for 3,308 positives) while IBG must approach querying the full dataset to obtain comparable results. This strongly validates the hypothesis. GBLE's global propagation enables discovering positives multiple hops away from seeds, while IBG's local expansion gets stuck after exhausting immediate neighborhoods. The dynamic batch sizing in GBLE also helps: when precision drops (fewer positives nearby), GBLE automatically expands search radius. Among IBG variants, large $d_{\max}$ performs best and low values severely limits reachability. This suggests an optimal "Goldilocks zone" for local expansion, but global methods funda-

Dataset	All	Pos.	Neg.
Original SST2 Train	67349	37569	29780
Original Synthetic	5000	2488	2512
Imbalanced SST2 Train	33088	3308	29780
Synthetic Seeds (Positive)	100	100	0

Table 1: Dataset statistics for SST2.

Dataset	All	Pos.	Neg.
MHS Train	39565	2598	36967
Synthetic	1000	19	981
Synthetic Seeds (Positive)	19	19	0

Table 2: Dataset statistics for MHS.

mentally outperform. Based on these results, we select ACS seeds + GBLE as the optimal configuration for Phase 2 realistic evaluation.

5.4 Realistic Evaluation on MHS

Having validated design choices on controlled SST2, we now evaluate SYNAPSE-G on a naturally occurring rare event: detecting hate speech targeting transgender individuals in social media comments.

Dataset: The Measuring Hate Speech (MHS) dataset (Kennedy et al., 2020; Sachdeva et al., 2022) contains 39,565 social media comments (YouTube, Reddit, Twitter) with fine-grained hate speech annotations from 7,912 annotators. Each comment is labeled across 10 ordinal dimensions and tagged with targeted demographic groups.

We define a binary rare event label: a comment is positive if any annotator marked it as targeting transgender men, transgender women, or unspecified transgender individuals (logical OR across subcategories). This yields 2,598 positive comments (6.5%). We use 1,000 synthetic comments generated by Llama-2 (Casula et al., 2024) for hate speech topics. Of these, only 19 were positive for our transgender target. These 19 synthetic positives are all used as seeds (no selection needed given scarcity).

Baselines: We further design a practical iterative baseline, "LR-Baseline", which adopts an iterative active learning approach using a simple classifier. We establish the baseline using a logistic regression model trained on an initial set comprising 19 positive synthetic seeds augmented with 19 randomly sampled known negative instances from the dataset. These examples are represented using pre-trained Gecko embeddings. The iterative refinement process then proceeds as follows: In each iteration, a subset of unlabeled data points, constrained by an inference budget (B) to ensure practical feasibility by avoiding inference over the entire dataset, is selected. The current logistic regression model predicts positivity probabilities for this subset, and the top K candidates with the highest predicted probabilities are chosen for labeling via an oracle (K is dynamically adjusted as $K = K_0/p_{prev}$, where $K_0 = 100$, consistent with SYNAPSE-G). These newly labeled instances are then incorporated into the training set, and the logistic regression model is retrained. The inference budget (B) is a critical parameter for maintaining the practicality of the approach. We vary inference budget $B \in \{1000, 4000, 8000, 16000, 39565\}$ to explore computational cost vs. performance trade-off. Despite LR-Baseline's initial advantage (knowing true negatives), we show SYNAPSE-G achieves superior label efficiency.

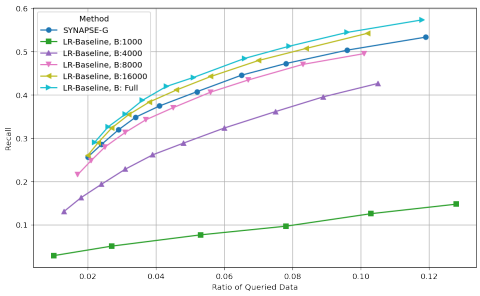


Figure 4: Recall vs. Ratio of Queried Data for *iterative* rare event detection on the imbalanced MHS dataset. Synapse-G (ACS seed selection + GBLE propagation) maintains a higher recall even when LR-baseline is endowed with a large inference budget.

Figure 4 compares recall vs. query ratio for SYNAPSE-G and LR-Baseline variants. With only 2.4% labeled (950 instances), SYNAPSE-G discovers 28.6% of positives, and with 5.0% labeled (1979 instances) it discovers 40.8% of positives. In contrast, LR-Baseline requires a large inference budget ($B = 8000$, inferring on 20% of the data per round) to reach similar recalls. Furthermore, only with very large or unlimited budgets ($B = 16k$ or Full), involving significant computational cost per iteration, does LR-Baseline outperform SYNAPSE-G in terms of recall. For context, at billions of instances (e.g., social media monitoring), inferring on 20% per iteration means processing 200M instances per round, which is orders of magnitude more than SYNAPSE-G's one-time graph construction. Observe that the performance gap is largest at low labeling budgets (where rare event scenarios operate). This demonstrates SYNAPSE-G's design specifically targets the cold-start regime where model-based active learning struggles. As more labels accumulate, LR-Baseline improves, but SYNAPSE-G maintains its lead through superior initial discovery.

REFERENCES

Shumeet Baluja, Rohan Seth, Dharshi Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pp. 895–904, 2008.

Richard A Bauder and Taghi M Khoshgoftaar. A study on rare fraud predictions with big medicare claims fraud data. *Intelligent Data Analysis*, 24(1):141–161, 2020.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760, 2020.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. *Semi-Supervised Learning*, pp. 193–216, 2006.

Ander Carreño, Iñaki Inza, and Jose A Lozano. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53: 3575–3594, 2020.

Camilla Casula, Sebastiano Vecellio Salto, Alan Ramponi, Sara Tonelli, et al. Delving into qualitative implications of synthetic data for hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19709–19726, 2024.

Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14855–14865, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.873. URL https://aclanthology.org/2024.findings-emnlp.873/.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is gpt-3 a good data annotator? In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1679–1705, 2024.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. Better synthetic data by retrieving and transforming existing datasets. *arXiv preprint arXiv:2404.14361*, 2024.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*, 2022.

Matthew Herland, Richard A Bauder, and Taghi M Khoshgoftaar. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *Journal of Big Data*, 6:1–33, 2019.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 113–122, 2021.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pp. 173–182, 2019.

Sujith Ravi and Qiming Diao. Large scale distributed semi-supervised learning using streaming approximation. In *Artificial intelligence and statistics*, pp. 519–528. PMLR, 2016.

Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pp. 232–241. Springer, 1994.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pp. 83–94, 2022.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International symposium on intelligent data analysis*, pp. 309–318. Springer, 2001.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles. Active learning literature survey. 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Chathurangi Shyalika, Ruwan Wickramarachchi, and Amit P Sheth. A comprehensive survey on rare event prediction. *ACM Computing Surveys*, 57(3):1–39, 2024.

Alexandra A Siegel. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pp. 56–88, 2020.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Victor Suarez-Lledo and Javier Alvarez-Galvez. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187, 2021.

Partha Talukdar and William Cohen. Scaling graph-based semi supervised learning to large number of labels using count-min sketch. In *Artificial Intelligence and Statistics*, pp. 940–947. PMLR, 2014.

Sasan Tavakkol, Max Springer, Mohammadhossein Bateni, Neslihan Bulut, Vincent Cohen-Addad, and MohammadTaghi Hajiaghayi. Less is more: Adaptive coverage for synthetic training data. *arXiv preprint arXiv:2504.14508*, 2025.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11653–11669, 2022.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *ProQuest number: information to all users*, 2002.

A Related Work

Our work bridges several research areas: active learning for rare events, synthetic data generation, retrieval methods, graph-based semi-supervised learning, and diversity sampling.

Active Learning and Rare Event Detection. Active learning addresses the challenge of training models with limited labeling budgets by intelligently selecting which instances to label (Settles, 2009). Standard query strategies include uncertainty-based methods which select instances where the current model is least confident, measured via prediction entropy (Lewis, 1995), margin sampling (Scheffer et al., 2001), or variation ratios.Query-by-committee approaches train multiple models and select instances where predictions disagree most (Seung et al., 1992). Expected model change strategies select instances that would most alter model parameters if labeled (Settles et al., 2007). Diversity-based methods like CoreSet (Sener & Savarese, 2017) select representative instances that cover the feature space.

However, these approaches share a critical limitation: they require an initial model trained on at least some labeled examples. In the cold-start scenario we address (where zero labeled instances of the rare event exist) these methods cannot bootstrap. One must either (1) randomly sample until finding enough positives to train an initial model, which is prohibitively expensive for rare events (e.g., finding 10-20 examples in a 0.1% prevalence dataset requires labeling 10,000-20,000 instances), or (2) use heuristics like keyword search, which may miss nuanced or coded expressions of the rare event.

SYNAPSE-G's innovation is using LLM-generated synthetic examples to bypass the cold-start problem entirely. Rather than relying on labeled real data, we use synthetic seeds to construct an initial representation of the rare event in embedding space, then leverage graph structure to discover real positives. This enables effective active learning from iteration one, without requiring preliminary random sampling.

Related to our setting, positive and unlabeled (PU) learning (Bekker & Davis, 2020) addresses scenarios with only positive labels and unlabeled data. However, PU learning assumes access to a set of real labeled positives, whereas we start with only synthetic positives and fully unlabeled real data. Our theoretical analysis characterizes how the validity and diversity of these synthetic seeds impact discovery performance.

Synthetic Data Generation and Quality. Recent work has demonstrated LLMs' capability to generate synthetic training data for various NLP tasks (**?**Liu et al., 2024; Gandhi et al., 2024). However, synthetic data exhibits well-documented limitations: redundancy, distribution mismatch with real data, and potential for amplifying biases (Liu et al., 2024).

Most prior work uses synthetic data for direct model training by generating large synthetic datasets, then fine-tuning classifiers on this data (**?**Ye et al., 2022). Recent efforts like AlpaGasus (Chen et al., 2023) improve this by filtering synthetic data, keeping only high-quality examples rated by an LLM. Adaptive Coverage Sampling (ACS) (Tavakkol et al., 2025) addresses redundancy by selecting diverse subsets that maximize coverage of the semantic space.

Our approach differs fundamentally. Rather than training models on synthetic data, we use it as search queries to discover real instances. This paradigm shift addresses synthetic data's limitations. Specifically, we don't require synthetic examples to perfectly match real distributions, only to be semantically proximate enough to guide discovery. Real labeled instances (verified by oracle) ultimately train the final classifier, avoiding distribution mismatch issues.

We leverage ACS for seed selection (Section 5) to maximize diversity, but our theoretical analysis reveals a nuanced trade-off: under certain conditions, excessive diversity can reduce precision (Proposition 4.1). This insight extends beyond prior work's focus on diversity as uniformly beneficial.

Retrieval and Semantic Search. Retrieval methods identify relevant items from large datasets given a query. Traditional lexical methods like BM25 (Robertson & Walker, 1994) rely on term frequency and inverse document frequency (TF-IDF), which fail to capture semantic similarity. Dense

retrieval methods (Karpukhin et al., 2020; Lee et al., 2019; Xiong et al., 2020; Izacard et al., 2021) embed queries and documents into shared vector spaces, enabling semantic matching.

Particularly relevant is Hypothetical Document Embeddings (HyDE) (Gao et al., 2022), which generates a synthetic document for a query, then retrieves real documents similar to this synthetic example. Thus, avoiding the need for relevance labels or task-specific fine-tuning.

SYNAPSE-G extends this idea from single-query retrieval to iterative discovery with graph structure. While HyDE generates one hypothetical document per query, we generate multiple diverse synthetic seeds and propagate their influence through a similarity graph to identify candidates. Furthermore, we analyze the theoretical properties of this process, characterizing how seed quality affects discovery performance (a question not addressed in retrieval literature).

Our work also connects to hard negative mining in contrastive learning (Xiong et al., 2020; Hofstätter et al., 2021), which selects challenging negatives near decision boundaries to improve model generalization. We adapt this intuition in reverse: using synthetic positives to mine real positives in extremely imbalanced settings.

Graph-Based Semi-Supervised Learning. Label propagation (Zhu & Ghahramani, 2002; Bengio et al., 2006) is a foundational semi-supervised learning technique that leverages graph structure, assuming similar nodes should have similar labels. This smoothness assumption enables inferring labels for unlabeled instances given a partially labeled graph (Talukdar & Cohen, 2014; Ravi & Diao, 2016; Baluja et al., 2008).

Standard label propagation requires an initial set of real labeled examples and assumes the graph structure accurately captures class boundaries. In the rare event cold-start scenario, we have neither: (1) no real labeled positives exist initially, and (2) the graph is dominated by negatives, making class boundaries difficult to identify without guidance.

SYNAPSE-G addresses this by using synthetic seeds to initialize the propagation process. The synthetic examples provide semantic anchors in embedding space, guiding the graph structure toward regions likely to contain real positives. Our Graph-Based Label Expansion (GBLE, Algorithm 2) adapts standard label propagation (Zhu & Ghahramani, 2002) by incorporating synthetic seeds as initial labels and using dynamic batch sizing based on precision feedback.

We also propose Iterative Bipartite Graph (IBG, Algorithm 1), a more conservative local expansion strategy. Our experiments (Section 5) demonstrate that GBLE's global approach outperforms IBG's local strategy in iterative discovery scenarios, providing empirical guidance for practitioners.

Overall our contributions synthesize insights from these areas to address a problem not fully solved by any single approach: efficiently discovering rare events in massive unlabeled datasets with zero initial real labels. The theoretical framework and empirical validation demonstrate that this synthesis yields practical benefits, substantially reducing labeling costs compared to standard active learning baselines.

B Algorithm Pseudocode

C Omitted Proofs

C.1 Proof of Proposition 4.1

We first reprint the proposition for readability.

Proposition C.1. *Let $Q \coloneqq S \cup N(S_+)$ denote the queried vertices and P be the number of positive examples in Q. Then,*

$$\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] = (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h\left(S_+\right)}$$

$$\mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\left((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\right)$$

Algorithm 1 Iterative Bipartite Graph (IBG)

Require: Unlabeled data $\mathcal{D}_U$, Initial positive seeds $\mathcal{D}_S$, Similarity threshold τ, Maximum degree $d_{\max}$, Number of iterations T
Ensure: Labeled data $\mathcal{L}$

Initialize $V_P = \mathcal{D}_S$, $\mathcal{L} = \{(x, 1) : x \in \mathcal{D}_S\}$ // Seeds assumed positive
for $t = 1$ **to** T **do**
  $V_U = \mathcal{D}_U \setminus \bigcup_{j=1}^{t-1} \mathcal{B}_j$ // Remaining unlabeled data
  Construct bipartite graph $G_B = (V_P, V_U, E_B)$ where:
    $E_B = \{(v_i, v_j) : v_i \in V_P, v_j \in V_U, \text{sim}(v_i, v_j) > \tau\}$
  for each $v_i \in V_P$ **do**
    $N_i \leftarrow \{v_j \in V_U : (v_i, v_j) \in E_B\}$ // Neighbors of v_i
    Sort N_i by $\text{sim}(v_i, \cdot)$ in descending order
    Keep only top $d_{\max}$ neighbors in N_i, remove others from E_B
  end for
  $\mathcal{B}_t \leftarrow \{v_j \in V_U : \exists v_i \in V_P, (v_i, v_j) \in E_B\}$ // Collect candidates
  Obtain labels $\mathcal{L}_t = \{(x, y) : x \in \mathcal{B}_t\}$ from oracle
  $V_P \leftarrow V_P \cup \{v \in \mathcal{B}_t : y(v) = 1\}$ // Add new positives
  $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_t$
  $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{B}_t$
end for
return $\mathcal{L}$

Algorithm 2 Graph-Based Label Expansion

Require: Similarity graph $G = (V, E)$, initial labels Y_0 (partially labeled), iterations T.
Ensure: Final label assignments Y_T.

Initialize $Y^{(0)} = Y_0$.
for $t = 1$ to T **do**
  Construct the normalized adjacency matrix W from G:

$$W_{ij} = \begin{cases} \frac{1}{\deg(v_i)} & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

  Propagate labels: $Y^{(t)} = WY^{(t-1)}$.
  Reinforce initial labels: For all nodes v_i with initial labels in Y_0, set $Y_i^{(t)} = Y_{0,i}$.
end for
return $Y^{(T)}$.

where $\frac{P}{|Q|}$ *is the precision and* $\frac{P}{|V|}$ *is the recall.*

Proof. $Q = S \cup N(S_+) = S_+ \cup S_- \cup N(S_+) = S_- \cup N(S_+)$. Since S is an independent set (Assumption 1), $S_- \cup N(S_+)$ is disjoint. Thus,

$$\begin{aligned} |Q| &= |S_-| + |N(S_+)| \\ &= (1-p)|S| + h(S_+)p|S| \\ &= (1 - p + ph(S_+))|S|. \end{aligned}$$

Let $S_1 \subseteq N(S_+) \setminus S_+$ be vertices in $N(S_+) \setminus S_+$ adjacent to exactly one vertex in S_+, and S_2 be those adjacent to exactly two. Let P_1, P_2 be the number of positive examples in S_1, S_2, respectively.

$$\begin{aligned} \mathbb{E}[P \mid S] &= \mathbb{E}[|S_+| + P_1 + P_2 \mid S] \\ &= |S_+| + q_1|S_1| + q_2|S_2|. \end{aligned}$$

We have:

$$|S_+| + |S_1| + |S_2| = |N(S_+)| \tag{1}$$
$$d|S_+| + |S_+| - |S_2| = |N(S_+)| \tag{2}$$
$$|N(S_+)| = |S_+|h(S_+). \tag{3}$$

Equation equation 1 counts vertices in $N(S_+)$. Equation equation 2 counts edges between S_+ and $N(S_+)$, subtracting $|S_2|$ once (as each is counted twice). Equation equation 3 is from the definition of $h(S_+)$. Solving equation 1-equation 3:

$$\begin{aligned}
|S_1| &= (2h(S_+) - d - 2)|S_+| \\
|S_2| &= (d + 1 - h(S_+))|S_+| \\
\mathbb{E}[P \mid S] &= |S_+|(1 + q_1(2h(S_+) - d - 2) \\
&\quad + q_2(d + 1 - h(S_+))) \\
&= p|S|((1 - 2q_1 + q_2) \\
&\quad + (q_2 - q_1)d + (2q_1 - q_2)h(S_+)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\
&= \frac{p|S|((1 - 2q_1 + q_2) + (q_2 - q_1)d + (2q_1 - q_2)h(S_+))}{(1 - p + ph(S_+))|S|} \\
&= (2q_1 - q_2) + \frac{1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right)}{\frac{1-p}{p} + h(S_+)}.
\end{aligned}$$

If $1 + q_2\left(d + \frac{1}{p}\right) - q_1\left(d + \frac{2}{p}\right) > 0$, then we must have that

$$p > \frac{2q_1 - q_2}{1 + (q_2 - q_1)d}$$

and $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ decreases with $h(S_+)$. Now, we compute the derivative of $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ with respect to p:

$$\begin{aligned}
&\frac{\partial}{\partial p}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\
&= \frac{(1 - q_1(2 + d) + q_2(1 + d)) + h(S_+)(2q_1 - q_2)}{(1 - p + ph(S_+))^2} \\
&= \frac{1 + d(q_2 - q_1) + (h(S_+) - 1)(2q_1 - q_2)}{(1 - p + ph(S_+))^2}.
\end{aligned}$$

Since $h(S_+) \leq d + 1$, $d \geq h(S_+) - 1$. Thus,

$$\begin{aligned}
&\frac{\partial}{\partial p}\mathbb{E}\left[\frac{P}{|Q|} \mid S\right] \\
&\geq \frac{1 + (h(S_+) - 1)(q_2 - q_1) + (h(S_+) - 1)(2q_1 - q_2)}{(1 - p + ph(S_+))^2} \\
&= \frac{1 + (h(S_+) - 1)q_1}{(1 - p + ph(S_+))^2} > 0.
\end{aligned}$$

Therefore, $\mathbb{E}\left[\frac{P}{|Q|} \mid S\right]$ is strictly increasing with respect to p.

Finally,

$$\begin{aligned}
\mathbb{E}\left[\frac{P}{|V|} \mid S\right] = \frac{p|S|}{|V|}\big((1 - 2q_1 + q_2) \\
+ (q_2 - q_1)d + (2q_1 - q_2)h(S_+)\big).
\end{aligned}$$

Since $q_2 < 2q_1$, $\mathbb{E}\left[\frac{P}{|V|} \mid S\right]$ increases with p and $h(S_+)$. □