# FRAGILITY-REVEALING GRAPH PERTURBATIONS FOR AMBIGUOUS SAMPLE DISCOVERY IN SHORT TEXT CLUSTERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Short text clustering is unstable when ambiguous samples are forced into hard clusters. Repeated clustering measures instability caused by algorithmic randomness. In contrast, we cast ambiguous sample discovery as a graph sparsification problem designed to reveal ambiguity. We propose the fragility-revealing graph perturbation (FRGP) method, an unsupervised method that aims to maximize the structural fragility of the semantic graph. We derive a semantic spanning subgraph that maximizes sensitivity to edge removal. Among such connected spanning subgraph, the minimum spanning tree has the least semantic distortion. Theoretically, under common assumptions on cluster separation and kNN connectivity, the entropy induced by these perturbations separates intra-cluster samples from structurally ambiguous samples. We further prove that the designed edge-weighted cutting mechanism gives a larger lower bound on the entropy gap than uniform cutting. Simulations verify the theoretical separation results. Experiments and ablation studies show that FRGP effectively identifies ambiguous samples, thereby making the retained clusters more compact for downstream use.

## 1 INTRODUCTION

Short text clustering is a fundamental problem in natural language processing (Ahmed et al., 2022). It aims to group unlabeled texts into coherent clusters without human annotation. This task supports a wide range of applications (Tussupov et al., 2025), including information retrieval (Schütze et al., 2008), recommendation systems (Aggarwal, 2016), social media analysis (Hettiarachchi et al., 2022), and user intent discovery (Shah et al., 2025); mining topics, creating user profiles, and recommending products likewise rely on clustering results (Blei, 2012).

Compared with long documents, short texts exhibit distinct challenges. They are extremely sparse and contain limited contextual information. Their semantics are often fragmented and incomplete. These properties make short text representations highly sensitive to small perturbations in the embedding space (Abdalgader et al., 2024). Even with modern neural encoders, minor changes in representations could lead to large variations in clustering assignments (Zhou et al., 2024). Consequently, short text clustering suffers from poor clustering stability, and the cluster boundaries are unreliable (Von Luxburg et al., 2010).

A key factor behind this clustering instability is the presence of samples with uncertain cluster membership. Such samples arise in two forms: inter-cluster points share semantic characteristics with multiple clusters (Ghaffari & Ghadiri, 2016), whereas globally out-of-distribution points have no clear membership in any cluster and lie far from all cluster cores (Sun et al., 2022). The latter can often be screened using distance- or confidence-based detectors (Hendrycks & Gimpel, 2017). We focus on the former and refer to them as ambiguous samples. These samples are not merely low-confidence points, but structurally fragile points that depend on weak semantic connections bridging multiple cluster cores. Their assignments fluctuate under controlled perturbations of the global semantic connectivity structure. Existing approaches characterize ambiguity through soft assignment scores or model disagreement (Ghaffari & Ghadiri, 2016; Ding et al., 2023), which capture local model uncertainty rather than this structural fragility. Therefore, identifying ambiguous samples requires a mechanism that explicitly exposes the fragility of the global semantic graph.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

**Our contributions.** (i) We formalize ambiguous sample discovery as estimating perturbation-induced cluster-assignment distributions. (ii) We prove that perturbation-induced entropy separates intra-cluster from ambiguous samples, with finite-sample bounds on the gap. (iii) Under matched-ablation conditions, we show that the bridge-sensitive backbone and distance-aware cutting improve the theoretical entropy-separation guarantee.

## 2 Related Work

### 2.1 Short text clustering

In text clustering, short texts contain only limited vocabulary and contextual cues. The fragmented semantics of short texts lead to unstable similarity measurements in the embedding space. Thus, clustering stability becomes a particularly prominent challenge in short text scenarios.

Existing methods could be categorized into three groups. The first group is representation learning and aims to improve clustering performance by learning high-quality embeddings (Li et al., 2024b), including multi-view representation refinement (Li et al., 2024b; Kamthawee et al., 2024), mutual-information regularization (Yao, 2025), and pseudo-label generation (Yao et al., 2025). These methods improve embedding separability, but they usually treat ambiguous samples as ordinary points rather than explicitly modeling semantic ambiguity.

The second group of methods adopts a network structural modeling perspective. Liu et al. (2025) construct similarity-based graphs to model short texts and apply graph-based clustering algorithms. Further refinement of the modeling involves capturing global relationships between samples using common information propagation patterns on the network (Vu et al., 2025). Furthermore, Wang et al. (2025a) utilize a large language model to generate pairwise or triplet constraints between samples, obtaining several subgraphs characterizing the local structure of the samples and clustering based on these. Existing methods based on subgraphs, such as minimum spanning trees (MSTs), focus on identifying irregularly shaped clusters (Zhou et al., 2015), and detecting global outliers (Li et al., 2024a). However, they do not explicitly model cluster boundaries and thus remain limited in addressing clustering instability.

In recent years, short text clustering methods assisted by large language models have attracted increasing attention (Wu et al., 2025). The injected auxiliary information takes several forms (Naveed et al., 2025). Lin et al. (2025) use large language models to generate supplementary topic-level semantic descriptions. The auxiliary supervision signals could guide representation toward more discriminative directions (Yang et al., 2024; Zou et al., 2025; Wang et al., 2025b). Beyond direct semantic enhancement, some studies explore the co-training of large language models and embedding models, enabling the embedding space to progressively adapt in clustering (Li et al., 2025).

### 2.2 Ambiguous sample identification

The ambiguity arises from sensitive semantic structures. Ambiguous samples refer to inter-cluster points with uncertain category membership, rather than globally out-of-distribution points. These samples share semantic characteristics with multiple clusters and may consequently receive comparable membership scores from more than one cluster (Ghaffari & Ghadiri, 2016). Identifying such samples is important for clustering and classification (Zhou et al., 2022; Ding et al., 2023).

Existing studies mainly characterize ambiguous samples through structural consistency. A representative assumption is that ambiguous samples exhibit inconsistent representations across different structural views or neighborhood scales (Jin et al., 2021). Based on this idea, multi-scale contrastive methods compare node behavior between local substructures and global contexts (Zheng et al., 2021; Zhang et al., 2022). Liu et al. (2023) further evaluates ambiguity across node-level, graph-level, and population-level representations, while Pan et al. (2023) emphasizes local structural consistency between a sample and its neighborhood. These methods aim to identify unstable samples from multiple structural perspectives (Duan et al., 2023).

Other related methods use complementary signals. Ding et al. (2019) combine structural consistency with node attribute reconstruction error. Wu et al. (2023) measure the deviation of a sample from the dominant data distribution. In parallel, mixed-membership models describe ambiguity by allowing a

sample to belong to multiple clusters with different probabilities rather than a hard assignment (Tian et al., 2023; Qing & Wang, 2024b;a; Qing, 2025). Entropy-based measures have also been used to quantify clustering uncertainty (Hoayek & Rullière, 2025).

However, prior methods do not explicitly model the role of ambiguous samples as structurally fragile points in the global semantic connectivity structure. Thus, FRGP formulates ambiguous sample discovery as an ambiguity-revealing graph sparsification problem. The following section develops this idea into a concrete methodology.

# 3 METHODOLOGY

## 3.1 FRAGILITY-REVEALING GRAPH PERTURBATION (FRGP) METHOD

In a dense or multi-path semantic graph, removing a weak boundary edge often has little effect on connectivity because alternative paths can bypass the perturbation, thereby masking the instability of ambiguous samples. We therefore seek a bridge-sensitive semantic backbone that preserves global connectivity while maximizing edge-removal sensitivity. Concretely, given a connected spanning subgraph $H = (\mathcal{V}, \mathcal{E}_H)$, $\mathcal{E}_H \subseteq \mathcal{E}_K$, we define the *edge-removal fragility*:

$$\Phi(H) \triangleq \frac{|\{e \in \mathcal{E}_H : e \text{ is a bridge of } H\}|}{|\mathcal{E}_H|}. \tag{1}$$

Intuitively, $\Phi(H)$ measures the fraction of edges whose removal would disconnect $H$.

**Definition 3.1** (Bridge-sensitive backbone)**.** The edge-removal fragility satisfies $\Phi(H) = 1$ if and only if $H$ is a spanning tree. The minimum spanning tree $T^* \in \arg\min_{T \in \mathcal{T}(G_K)} \sum_{(i,j)\in T} \delta_{ij}$ minimizes the semantic distortion.

The minimum semantic distortion ensures that the backbone faithfully preserves the original semantic structure, so that the subsequent perturbations reveal semantic fragility. Building on Definition 3.1, we propose the **fragility-revealing graph perturbation (FRGP)** method. The FRGP method consists of the following four steps, as illustrated in Figure 1.

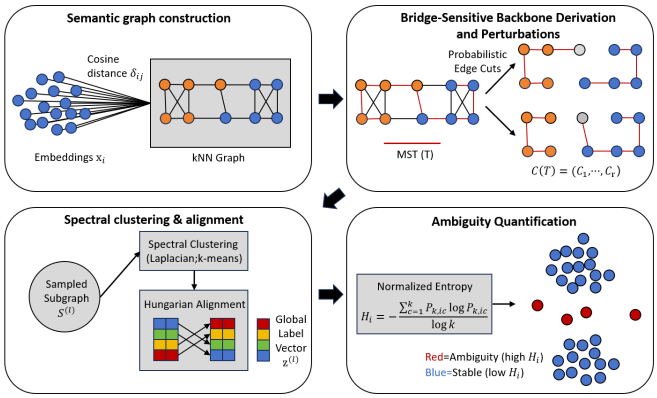


Figure 1: Overview of the FRGP method.

**Step 1: Semantic network construction.** This step converts short texts into a weighted semantic network, where nodes are texts and edges represent local semantic similarity. We encode each text as an embedding $x_i \in \mathbb{R}^d$ for $i \in \mathcal{V} = \{1, \ldots, n\}$, and measure semantic dissimilarity by the cosine distance $\delta_{ij} \triangleq 1 - \frac{x_i^\top x_j}{\|x_i\| \, \|x_j\|} \in [0, 2]$. The corresponding similarity weight is $w_{ij} \triangleq \min\{\max\{1 - \delta_{ij}, 0\}, 1\}$. Given the kNN parameter $K$, we define $\mathcal{N}_K(i) \triangleq \arg\min_{S \subseteq \mathcal{V}\setminus\{i\}, |S|=K} \sum_{j\in S} \delta_{ij}$ and construct the symmetric edge set $\mathcal{E}_K \triangleq \{(i,j) : j \in \mathcal{N}_K(i) \text{ or } i \in \mathcal{N}_K(j)\}$. The resulting weighted undirected semantic network is $G_K$.

**Step 2: Bridge-sensitive backbone derivation.** This step extracts a bridge-sensitive semantic backbone from the original semantic network and perturbs its weak links to expose assignment instability. Following the edge-removal fragility in equation 1 and the minimum-distortion characterization in Definition 3.1, we define the bridge-sensitive semantic backbone as $T \triangleq \mathrm{MST}(G_K)$. Since $G_K$ uses $w_{ij}$ as the similarity weight for spectral clustering and $\delta_{ij}$ as the edge distance for MST construction, $T$ maximizes $\Phi(H)$ among connected spanning subgraphs while minimizing total semantic distance among spanning trees, removing redundant kNN paths but preserving global connectivity. For each edge $(i,j) \in T$, let $d_{ij}^{\mathrm{MST}} \triangleq \delta_{ij}$. Longer edges are more likely to represent weak semantic links or boundary connections between clusters. To construct a deterministic reference partition for label alignment, we remove the $k-1$ edges with the largest cosine distances and take the resulting connected components as $\mathcal{C}(T) = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$. This perturbation makes ambiguous samples perturbation-sensitive while leaving stable intra-cluster samples less affected.

**Step 3: Probabilistic graph perturbation, spectral clustering, and alignment.** This step performs spectral clustering on the sampled subgraphs and aligns their labels so that the results from different perturbation rounds are comparable. We assign cutting probabilities $P = \{p_{ij} : (i,j) \in T\}$ with $p_{ij} = d_{ij}^{\mathrm{MST}} / \sum_{(u,v)\in T} d_{uv}^{\mathrm{MST}}$ for $(i,j) \in T$. We conduct $L$ independent perturbation rounds. At round $\ell$, we sequentially remove $k-1$ distinct edges from $T$ without replacement according to $P$, renormalized over the remaining edges after each cut, obtaining $\mathcal{C}^{(\ell)}(T) = \{\mathcal{C}_1^{(\ell)}, \ldots, \mathcal{C}_k^{(\ell)}\}$. One connected component is then selected uniformly from $\mathcal{C}^{(\ell)}(T)$, yielding a sampled subgraph $S^{(\ell)} \subseteq \mathcal{V}$. We apply spectral clustering to the subgraph induced by the sampled nodes on the original kNN graph $G_K$, which keeps the local semantic relationships while reducing the clustering scale. Following the standard spectral clustering pipeline, we compute the normalized network Laplacian and extract the smallest $r^{(\ell)}$ eigenvectors selected by the eigengap heuristic. We then apply $k$-means with $k$ centers to the resulting $r^{(\ell)}$-dimensional embedding. Here, $k$ is the number of semantic clusters, whereas $r^{(\ell)}$ is the number of spectral coordinates used at round $\ell$. Since cluster labels may be permuted across different rounds, we align them using the Hungarian algorithm. This produces a global labeling vector $\hat{z}^{(\ell)} \in \left(\{1,\ldots,k\} \cup \{\perp\}\right)^n$, where the reference partition is the fixed bridge-sensitive backbone-based clustering $\mathcal{C}(T)$ obtained by cutting $k-1$ edges of $T$, and nodes not included in $S^{(\ell)}$ are assigned a null label $\perp$.

**Step 4: Ambiguity quantification.** This step aggregates the aligned clustering results across perturbation rounds and uses normalized entropy to quantify each text's assignment instability. For each node $i \in \mathcal{V}$, let $L_i \triangleq \sum_{\ell=1}^{L} \mathbf{1}\{i \in S^{(\ell)}\}$ be its sampling count, and define its empirical assignment probability to cluster $c$ as $\hat{\pi}_{ic} \triangleq \frac{1}{L_i} \sum_{\ell=1}^{L} \mathbf{1}\{i \in S^{(\ell)}, \hat{z}_i^{(\ell)} = c\}$. This gives $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{ik}) \in \Delta^{k-1}$. We then define the normalized entropy of node $i$ as $\hat{H}_i \triangleq -\frac{1}{\log k} \sum_{c=1}^{k} \hat{\pi}_{ic} \log \hat{\pi}_{ic} \in [0,1]$. Intuitively, a low entropy means that the text is consistently assigned to the same cluster, while a larger $\hat{H}_i$ indicates more frequent assignment changes under perturbations, so it measures perturbation-induced instability rather than generic soft-assignment uncertainty. The entropy scores define a continuous ambiguity ranking.

**Application.** FRGP first produces a continuous ambiguity score and a descending ranking for each text. The ranking can be thresholded according to downstream requirements; the method does not assume a known ambiguous-sample proportion. Low-ambiguity samples can then train a lightweight model for online clustering, while the scores remain available for selective review.

### 3.2 Time and space complexity

Compared with dense full-graph spectral clustering, FRGP offers clear advantages in computational and memory efficiency. The detailed derivation is provided in Appendix F.

**Time complexity.** Time complexity of FRGP is $O\big(n^2 d + nK \log n + L\left(n + tk\bar{m} + \bar{s}k^2 I + k^3\right)\big)$, where $n$ is the sample size, $d$ is the embedding dimension, $K$ is the kNN parameter, $k$ is the number of clusters, and $L$ is the number of perturbation rounds. Here, $\bar{s}$ and $\bar{m}$ are the average sampled-subgraph size and edge count, while $t$ and $I$ are the eigensolver and k-means iteration counts.

**Algorithm 1:** Fragility-Revealing Graph Perturbation (FRGP) method

**Input :** Short-text embeddings $\{x_i\}_{i=1}^n$, embedding dimension $d$, number of clusters $k$, kNN parameter $K$, and subgraph sampling rounds $L$.

**Output:** The ranked entropy scores $\{\hat{H}_i\}_{i=1}^n$, together with cluster assignments $\{\hat{z}_i\}_{i=1}^n$.

**Step 1. Semantic network construction.** Compute cosine distance $\delta_{ij}$ for embeddings. Compute semantic similarity weight $w_{ij}$. Construct the symmetric kNN network $G_K$.

**Step 2. Derive bridge-sensitive semantic backbone.**

  Derive the bridge-sensitive semantic backbone of $G_K$: $T \triangleq \mathrm{MST}(G_K)$;

  Compute new distance $d_{ij}^{\mathrm{MST}}$ based on the subgraph $T$;

  Generate the fixed reference partition $\mathcal{C}(T)$ by removing the $k-1$ MST edges with the largest cosine distances;

**Step 3. Perturb probabilistic graph, run spectral clustering, then align.**

  Compute the cutting probabilities $p_{ij}$, forming the distance-aware cutting distribution $P$;

  **for** $\ell = 1$ **to** $L$ **do**

    Perform $k-1$ probabilistic edge-cuts on $T$ guided by $P$, obtaining $\mathcal{C}^{(\ell)}(T)$;

    Uniformly sample one connected component $S^{(\ell)}$ from $\mathcal{C}^{(\ell)}(T)$;

    Apply spectral clustering on $S^{(\ell)}$;

    Align cluster labels via the Hungarian algorithm.

  **end**

**Step 4. Ambiguity quantification.**

  Compute the normalized entropy scores $\hat{H}_i$. Return the cluster assignments $\{\hat{z}_i\}_{i=1}^n$.

**Space complexity.** With a sparse kNN graph and streaming aggregation across rounds, FRGP requires $O(nd + nK + nk)$ space for embeddings, the sparse graph, and per-sample assignment counters, without storing all $L$ clustering outputs.

**Comparison with dense full-graph spectral clustering.** Dense full-graph spectral clustering typically requires $O(n^2 d + n^3)$ time and $O(nd + n^2)$ space. FRGP reduces these costs by using kNN/MST sparsification and spectral clustering on sampled subgraphs, whose per-round cost is much smaller when $\bar{s} \ll n$.

## 4 THEORETICAL PROPERTIES

To establish the theoretical properties of FRGP, we impose the following assumptions. Throughout, $k \geq 2$ is fixed, $G_K$ is connected, and perturbation rounds are independent. For $i \in \mathcal{F}$ and $r \in \{a, b\}$, let $\mathcal{T}_{r(i)}$ denote a nearby semantic cluster, $B_i^r \subseteq E(T)$ its boundary-switching MST edge set, and $\Phi_i^r(H)$ the probability that a matched boundary perturbation changes the $i$–$\mathcal{T}_{r(i)}$ connectivity in graph $H$.

**Assumption 4.1** (Strongly separated semantic clusters)**.** The sample set $\mathcal{V}$ is partitioned into an intra-cluster set $\mathcal{T}$ and a complementary set of ambiguous samples $\mathcal{F}$, i.e. $\mathcal{V} = \mathcal{T} \cup \mathcal{F}$, $\mathcal{T} \cap \mathcal{F} = \varnothing$. The intra-cluster set $\mathcal{T}$ further consists of $k$ mutually disjoint tight clusters, $\mathcal{T} = \bigcup_{r=1}^k \mathcal{T}_r$, $\mathcal{T}_r \cap \mathcal{T}_s = \varnothing \quad (r \neq s)$. There exist constants $0 < \Delta_{\mathrm{in}} < \Delta_{\mathrm{out}}$ such that $\max_r \ \max_{i,j \in \mathcal{T}_r} \delta_{ij} \leq \Delta_{\mathrm{in}}, \min_{r \neq s} \ \min_{i \in \mathcal{T}_r,\, j \in \mathcal{T}_s} \delta_{ij} \geq \Delta_{\mathrm{out}}$.

**Assumption 4.2** (Connectivity preservation by kNN)**.** For each $r \in \{1, \ldots, k\}$, the subgraph of the $k$-nearest-neighbor graph $G_K$ induced on $\mathcal{T}_r$ is connected.

**Assumption 4.3** (Structurally ambiguous samples)**.** Let $\Delta_{\mathrm{in}} < \Delta_{\mathrm{ambiguous}} \leq \Delta_{\mathrm{near}} < \Delta_{\mathrm{out}}$. For any $i \in \mathcal{F}$, $\min_{j \in \mathcal{T}_r} \delta_{ij} \geq \Delta_{\mathrm{ambiguous}} \quad \forall r$, and there exist distinct $\mathcal{T}_{a(i)}$ and $\mathcal{T}_{b(i)}$ such that $\min_{j \in \mathcal{T}_{a(i)}} \delta_{ij}, \ \min_{j \in \mathcal{T}_{b(i)}} \delta_{ij} \leq \Delta_{\mathrm{near}}$. For each $r \in \{a, b\}$, there exists a nonempty boundary-switching set $B_i^r \subseteq E(T)$ such that removing any $e \in B_i^r$ changes the $i$–$\mathcal{T}_{r(i)}$ connectivity, and, on a successful sampled round, activating $B_i^r$ alone induces label $r(i)$. These boundary edges are, on average, longer than the MST edges overall.

**Assumption 4.4** (Coverage of random subgraph sampling)**.** The random subgraphs $\{S^{(\ell)}\}_{\ell=1}^{L}$ are generated independently across rounds. There exists a constant $p_{\min} > 0$ such that $\mathbb{P}(i \in S^{(\ell)}) \geq p_{\min}(i \in \mathcal{V},\ \ell \in [L])$.

**Assumption 4.5** (Correctness of alignment)**.** Let $A^{(\ell)}$ denote the event that spectral clustering and Hungarian alignment at round $\ell$ are correct. There exists $\varepsilon_n \to 0$ such that $\Pr\big((A^{(\ell)})^c \mid i \in S^{(\ell)}\big) \leq \varepsilon_n, i \in \mathcal{V},\ \ell \in [L]$. On $A^{(\ell)}$, every sampled $i \in \mathcal{T}_r$ receives semantic label $r$.

Assumptions 4.1 and 4.2 are common in theoretical analyses of clustering. Assumption 4.1 imposes within-cluster compactness and between-cluster separation (Ng et al., 2001; Schiebinger et al., 2015), while Assumption 4.2 ensures within-cluster connectivity in the kNN graph (von Luxburg, 2007). Under these conditions, our guarantee targets structurally ambiguous samples. Specifically, these samples lie near multiple cluster cores and depend on weak semantic bridges. Assumption 4.3 characterizes ambiguous samples as points separated from cluster interiors but close to multiple clusters, with relatively long boundary-related connections on the MST. Assumption 4.4 ensures persistent sampling coverage across independent perturbation rounds. Assumption 4.5 follows standard spectral-clustering theory: spectral clustering can recover the underlying cluster structure up to label permutation with high probability (Lei & Rinaldo, 2015). The remaining label permutation can then be resolved by Hungarian matching (Kuhn, 1955).

To characterize the effective boundary switching induced by FRGP, MST and distance-aware cutting, at round $\ell$, we let $C^{(\ell)}$ be the removed MST edges and define $E_i^{r,(\ell)} = \{C^{(\ell)} \cap B_i^r \neq \varnothing,\ C^{(\ell)} \cap B_i^t = \varnothing\}, t \neq r$, which means that only the boundary toward cluster $r(i)$ is activated. Finally, $q_i^{\mathrm{FRGP}} = \min_{r \in \{a,b\},\, \ell \in [L]} P\Big(E_i^{r,(\ell)} \mid i \in S^{(\ell)}\Big) - \varepsilon_n$, where $S^{(\ell)}$ is the sampled subgraph and $\varepsilon_n$ is the alignment error bound. Thus, $q_i^{\mathrm{FRGP}}$ measures the minimum effective boundary-switching probability of sample $i$.

**Proposition 4.6** (MST maximizes perturbation effectiveness)**.** *For any connected graph $H$, uniformly deleting one edge disconnects $H$ with probability $\Phi(H)$. For $T = \mathrm{MST}(G_K)$, $\Phi(T) = 1$, whereas $\Phi(G_K) < 1$ whenever $G_K$ contains a cycle. For an ambiguous sample $i \in \mathcal{F}$ and $r \in \{a, b\}$, if a positive-probability subset of the boundary perturbations can be bypassed by alternative paths in $G_K$, then $\Phi_i^r(T) = 1 > \Phi_i^r(G_K)$.*

**Proposition 4.7** (Distance-aware cutting favors ambiguous boundary edges)**.** *Under Assumption 4.3, for every ambiguous sample $i \in \mathcal{F}$ and $r \in \{a, b\}$, distance-aware MST cutting assigns a larger probability to the boundary-switching set $B_i^r$ than uniform MST cutting: $P_{\mathrm{FRGP}}(B_i^r) > P_U(B_i^r)$. For sequential cutting, the same ordering holds as long as the boundary set remains relatively overweight among the remaining edges and is not hit with probability one.*

Proposition 4.6 shows that the MST removes redundant kNN edges, making every edge a bridge, so that cutting alters the relative connectivity between ambiguous samples and nearby clusters; this raises the effective boundary-switching probability compared with the original graph. Proposition 4.7 then shows that, because ambiguous samples are connected through longer and weaker semantic links, distance-aware cutting assigns higher probability to their boundary edges than uniform cutting. Under the matched-ablation conditions in Appendix B.6, these two effects combine into the lower-bound advantage of Corollary 4.9. Consequently, ambiguous samples exhibit less stable assignments across perturbation rounds and hence higher label entropy, whereas intra-cluster samples remain more stable within their clusters.

**Theorem 4.8** (Entropy separability induced by FRGP perturbations)**.** *Under Assumptions 4.1–4.5, suppose $\min_{i \in \mathcal{V}} L_i / \log n \xrightarrow{p} \infty$, $q_*^{\mathrm{FRGP}} := liminf_{n \to \infty} inf_{i \in \mathcal{F}} q_i^{\mathrm{FRGP}} > 0$. Then there exist constants $0 < h_{\mathrm{in}} < h_{\mathrm{amb}} \leq 1$ and $\tau \in (h_{\mathrm{in}}, h_{\mathrm{amb}})$ such that*

$$\mathbb{P}\left(\max_{i \in \mathcal{T}} \widehat{H}_i \leq \tau,\ \min_{i \in \mathcal{F}} \widehat{H}_i \geq \tau\right) \to 1. \tag{2}$$

*Moreover, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $i \in \mathcal{V}$,*

$$\widehat{H}_i \leq g(2\varepsilon_n) + g(r_i(\delta)), i \in \mathcal{T}, \tag{3}$$

$$\widehat{H}_i \geq \frac{h_2(q_i^{\mathrm{FRGP}})}{\log k} - g(r_i(\delta)), i \in \mathcal{F}, \tag{4}$$

*where* $r_i(\delta) = k\sqrt{\frac{\log(2nk/\delta)}{2L_i}}$, $g(t) = \frac{h_2(t/2)+(t/2)\log(k-1)}{\log k}$, $h_2(u) = -u\log u-(1-u)\log(1-u)$, *whenever* $r_i(\delta) \le 2(1-1/k)$ *and* $\varepsilon_n \le 1-1/k$.

**Corollary 4.9** (Advantage in the entropy-separation lower bound)**.** *For* $M \in \{\mathrm{FRGP}, \mathrm{A1}, \mathrm{A2}\}$, *define* $q^{(M)}_{\min,n} = \min_{i\in\mathcal{F}} q^{(M)}_i$, $\Gamma_M = h_2(q^{(M)}_{\min,n})/\log k - g(2\varepsilon_n)$, *where* $q^{(M)}_i$ *is defined analogously to* $q^{\mathrm{FRGP}}_i$ *for mechanism* $M$. *Here A1 denotes direct sampling on* $G_K$ *without the MST, and A2 denotes uniform cutting on the same MST. Suppose* $0 < q^{(M)}_{\min,n} \le 1/2$ *for all* $M$. *Under the matched-ablation conditions specified in Appendix B.6 and the conditions of Propositions 4.6–4.7,*

$$\Gamma_{\mathrm{FRGP}} > \max\{\Gamma_{\mathrm{A1}}, \Gamma_{\mathrm{A2}}\}. \tag{5}$$

The proofs are given in Appendix B. Theorem 4.8 shows that FRGP perturbations induce separable assignment entropy: intra-cluster samples become low-entropy, whereas ambiguous samples retain non-vanishing entropy due to their positive effective boundary-switching probabilities. The separation is governed by three quantities: the clustering and alignment error, the effective boundary-switching probability, and the sampling count. The entropy gap widens as the clustering and alignment error decreases, as the effective boundary-switching probability increases, and as the sampling count grows. These dependences are explicit: the upper bound for intra-cluster samples decays with the error and the sampling count, while the lower bound for ambiguous samples is anchored by the boundary-switching probability. Corollary 4.9 shows that, by combining MST-based structural perturbation with distance-aware cutting, FRGP yields a larger lower bound on population entropy separation than either ablation. Simulation studies are reported in Appendix C. They verify entropy separation properties under varying sample sizes, ambiguity levels, and cluster separations.

# 5 Real application

## 5.1 Datasets

The real-world dataset is collected from a major Chinese food delivery platform during March–December 2024. It contains about 28,000 menu items from 1,500 restaurants across five universities, mostly with dish names shorter than ten Chinese characters. We use DeepSeek V3.2 (`https://huggingface.co/deepseek-ai/DeepSeek-V3.2`) to generate semantic descriptions and then encode them into 1792-dimensional vectors using the acge-text-embedding model (`https://huggingface.co/aspire/acge_text_embedding`). Ambiguous samples mainly correspond to bundled meals, multi-category dishes, or items with auxiliary descriptions, accounting for more than $15\%$ of the data. The ambiguity labels are derived from the platform's category-operating rules. This dataset provides a representative testbed for large-scale real-world short-text clustering. Overall, the data scenario exhibits an "intra-cluster samples + ambiguous samples" structure: most dishes belong to clear categories, while composite items concentrate near cluster boundaries.

We further evaluate FRGP on three public short-text benchmarks, including NLU++, a multi-intent dataset for which we pool all 20 official folds from the BANKING and HOTELS domains (Casanueva et al., 2022). AG News and StackOverflow are standard short-text clustering benchmarks containing 8,000 and 20,000 samples across 4 and 20 classes, respectively (Rakib et al., 2020). We encode all texts into 1,024-dimensional embeddings using BGE-M3 (`https://huggingface.co/BAAI/bge-m3`). Reference ambiguity labels are derived from the consistency of the original category annotations in indicating multi-cluster characteristics. These labels are used only for evaluation and are never provided to FRGP. For hyperparameter selection in each dataset, $k$ is selected from the eigengap of standard spectral clustering on a held-out validation subset. $K$ is set to the smallest value for which the corresponding kNN graph is connected. These settings are fixed before test evaluation.

## 5.2 Clustering performance comparison

**Baseline selection.** We compare FRGP with mixed-membership methods, including MMDF (Qing & Wang, 2024b), WDCMM (Qing, 2025), and ORC-L (Tian et al., 2023), which model soft or overlapping cluster structures through fractional memberships. We also include representative graph-based anomaly detection methods, including CoLA (Liu et al., 2022), ANEMONE (Jin et al., 2021),

Table 1: Comparison of label-free clustering quality (SS, ICS) and runtime on the real-world dataset.

| Method | FRGP | CoLA | ANEMONE | GRADATE | LLM |
|---|---|---|---|---|---|
| SS | **0.1114** | 0.0463 | 0.0242 | 0.0123 | 0.0403 |
| ICS | **0.5800** | 0.5099 | 0.5115 | 0.5159 | 0.5238 |
| Runtime | 20min57s | 59min28s | 3h13min | 6h12min | 31min24s |
| Method | SL-GAD | Sub | MMDF | WDCMM | ORC-L |
| SS | 0.0103 | 0.0037 | 0.0096 | 0.0080 | 0.0212 |
| ICS | 0.5507 | 0.5514 | 0.5159 | 0.5149 | 0.5093 |
| Runtime | 24min56s | 43min02s | 3min47s | 4min13s | 18h35min |

Table 2: Ambiguous-sample identification and clustering performance on public benchmarks.

| | AG News | | | StackOverflow | | | NLU++ | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | SS | ICS | AUC | SS | ICS | AUC | SS | ICS |
| FRGP | **0.7813** | **0.0971** | **0.5035** | **0.7271** | **0.1506** | **0.5898** | **0.8125** | **0.1441** | 0.6485 |
| MMDF | 0.7516 | 0.0341 | 0.4346 | 0.5873 | 0.0963 | 0.5233 | 0.6236 | 0.0952 | 0.6965 |
| WDCMM | 0.7371 | 0.0294 | 0.4193 | 0.5902 | 0.0965 | 0.5093 | 0.6309 | 0.1075 | 0.5001 |
| ORC-L | 0.6982 | 0.0603 | 0.4016 | 0.5827 | 0.1117 | 0.4485 | 0.5778 | 0.1153 | 0.6398 |

GRADATE (Duan et al., 2023), SL-GAD (Zheng et al., 2021), and Sub (Zhang et al., 2022). These methods cover contrastive, reconstruction-based, and multi-view self-supervised frameworks, and are widely used as strong graph-based baselines. In addition, DeepSeek V3.2 is used as an LLM baseline for direct ambiguous sample identification. For the real-world comparison in Table 1, all methods are evaluated on the same semantic embeddings derived from the real short-text data.

**Metric selection.** For datasets with reference ambiguity labels, we use AUC and Precision as the primary metrics for ambiguity discovery. SS and ICS are downstream metrics computed after applying the ambiguity ranking to obtain a retained sample set. SS measures intra-cluster cohesion and inter-cluster separation, with $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ and $\mathrm{SS} = \frac{1}{n}\sum_{i=1}^{n} s(i)$, where $a(i)$ and $b(i)$ are the within-cluster and nearest-cluster distances. ICS measures semantic compactness, defined as $\mathrm{IntraSim}(C_k) = \frac{1}{|C_k|(|C_k|-1)}\sum_{i\neq j,\, i,j\in C_k}\cos(x_i, x_j)$, and ICS is their cluster-size-weighted average. Higher SS and ICS indicate greater compactness and separation among the retained clusters. To make SS and ICS comparable, all methods remove the same proportion of top-ranked samples under ambiguity scores. Because the real-world ambiguity labels are derived from platform operating rules, Table 1 reports only the label-free metrics SS and ICS. AUC and Precision are reported in the ablation and public-benchmark evaluations, where reference ambiguity labels are available.

**Comparison with existing methods.** Table 1 shows that FRGP achieves the highest SS and ICS on the real-world dataset, indicating that filtering according to its ambiguity ranking yields more compact retained clusters. Existing methods are less effective because they mainly target isolated anomalies or soft memberships rather than boundary instability: CoLA and ANEMONE rely on node–subgraph contrastive signals, while SL-GAD, GRADATE, and Sub focus on attribute inconsistency or global–local feature fusion, which are insufficient for ambiguity caused by multiple cluster memberships. Mixed-membership methods such as MMDF, WDCMM, and ORC-L model soft memberships but tend to smooth boundary-specific instability. In contrast, FRGP amplifies perturbation-induced assignment fluctuations, making it more sensitive to structural ambiguity. We further compare FRGP with an LLM-assisted baseline in Table 1, where DeepSeek V3.2 uses an instruction prompt to assign each dish an ambiguity score from 1 to 10 (see Appendix D). FRGP achieves higher SS and ICS with lower runtime. This is because LLMs can recognize semantic ambiguity but do not model perturbation-sensitive boundary instability, whereas FRGP explicitly amplifies it through bridge-sensitive backbone perturbations.

**Performance on public benchmarks.** We evaluate ambiguous sample identification using the area under the receiver operating characteristic curve (AUC), computed from each method's continuous ambiguity scores. Table 2 shows that FRGP achieves the highest AUC on all three datasets, ranging from 0.7271 to 0.8125. It also obtains the highest SS on all three datasets and the highest ICS on AG News and StackOverflow, while its ICS ranks second to MMDF on NLU++. These results demon-

Table 3: Ablation study results on ambiguous sample identification and clustering performance.

| Method | Precision | | | AUC | SS | ICS |
|---|---|---|---|---|---|---|
| | 15% | 10% | 5% | | | |
| M | – | – | – | – | 0.0553 | 0.5077 |
| FRGP | **0.7625** | 0.6875 | 0.5625 | **0.8569** | **0.1114** | **0.5800** |
| A1 | 0.5541 | 0.7562 | 0.8500 | 0.7144 | 0.0425 | 0.5598 |
| A2 | 0.4708 | 0.3625 | 0.4500 | 0.8089 | 0.0698 | 0.5718 |
| A3 | 0.5833 | 0.4375 | 0.3125 | 0.8315 | 0.0410 | 0.5685 |

strate that FRGP consistently identifies ambiguous samples and generally improves the compactness of retained clusters across both multi-intent and conventional short-text clustering settings.

### 5.3 Ablation study

We conduct ablation studies to evaluate the roles of the graph sparsification structure, edge-weight-aware probabilistic perturbation, and the uniqueness of single-path connectivity in FRGP. Table 3 reports their effects on ambiguous sample identification and clustering performance, where **M** denotes clustering without ambiguous sample identification.

**Retaining ambiguous samples for clustering (M).** As shown in Table 3, FRGP yields higher SS and ICS on the retained samples than the no-filtering model M. These metrics quantify downstream compactness after filtering; they should be interpreted together with AUC and Precision for ambiguity discovery. M retains ambiguous samples, which blur cluster boundaries and reduce intra-cluster consistency, whereas FRGP removes ambiguous samples via bridge-sensitive backbone perturbation and edge-weight-aware cutting, improving cluster compactness and separation.

**Removing the graph sparsification, sampling directly on the network (A1).** In the A1 ablation, we remove the graph sparsification structure (i.e., the derived MST) and perform random sampling directly on the kNN graph. Redundant paths in the original kNN graph weaken the effect of edge removals on global connectivity, making the assignment instability of inter-cluster ambiguous samples less distinguishable from that of intra-cluster samples. As shown in Table 3, this causes clear drops in AUC, SS, and ICS, confirming that MST is crucial for suppressing redundant paths and amplifying boundary instability. Thus, directly perturbing the unsparsified kNN graph cannot reproduce the bridge-sensitive behavior of the MST.

**Retaining the graph sparsification, but with uniform edge cutting (A2).** A2 keeps the MST but replaces edge-weight-aware probabilistic cutting with uniform random cutting. Table 3 shows that A2 lowers Precision at all three ambiguous-sample ratios, as well as AUC, SS, and ICS, indicating that FRGP benefits from both the MST structure and distance-aware cutting. By preferentially cutting longer edges, FRGP exposes inter-cluster boundary uncertainty, whereas uniform cutting cannot distinguish weak boundary links from stable intra-cluster connections.

**Replacing single-path MST with multi-path structures (A3).** A3 replaces the MST's single-path connectivity with multi-path structures. As shown in Table 3, all metrics decline, indicating that alternative routes bypass cut edges and dilute perturbation-induced ambiguity signals. This confirms that the MST's unique-path property makes boundary assignments perturbation-sensitive.

## 6 Conclusion

This paper studies structural fragility of semantic similarity graphs in short-text clustering. We derive the bridge-sensitive semantic backbone as the spanning tree minimizing total semantic distance among all maximally fragile backbones. Under the stated conditions, our analysis establishes entropy-separation bounds for structurally ambiguous samples.

Several directions merit further investigation. First, the FRGP method can be extended to dynamic clustering settings. Second, broader evaluations should cover short-text benchmarks across different domains, languages, and ambiguity types. Third, ambiguity-aware identification could be integrated into downstream tasks, such as classification and recommendation.

## AI USE STATEMENT

In preparing this manuscript, the authors used generative AI tools only for translation and language polishing. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript. Separately, as an explicitly disclosed component of the experimental pipeline, DeepSeek was used to generate semantic descriptions and as an LLM baseline, as described in Sections 5.1- 5.2.

## ETHICS STATEMENT

This work studies unsupervised short-text clustering and ambiguous-sample identification. The real-world corpus contains menu-item text from a food-delivery platform and does not include personal or sensitive user information. The study did not involve human subjects and did not evaluate or intervene with individuals. The proprietary corpus will not be released, while the public benchmarks remain governed by their original licenses and terms of use. Potential risks include errors or biases inherited from the language and embedding models, which may affect downstream category management; therefore, model outputs should be reviewed before consequential deployment. We do not anticipate other direct societal harms arising from this work.

## REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe the FRGP procedure; Section 4 and Appendix B present the theoretical analysis and proofs; and Appendix C reports the simulation studies. Section 5 and Appendix D document the datasets, preprocessing pipeline, semantic-description and ambiguity-scoring prompts, public-benchmark construction, hyperparameter selection, and evaluation protocol. Appendix F provides the time and space complexity derivation. All experiments were repeated with Bootstrap resampling, and the conclusions were consistent across runs. The implementation and experiment configurations will be released after acceptance.

## REFERENCES


Khaled Abdalgader, Atheer A Matroud, and Khaled Hossin. Experimental study on short-text clustering using transformer-based semantic similarity measure. *PeerJ Computer Science*, 10:e2078, 2024.

Charu C Aggarwal. An introduction to recommender systems. In *Recommender systems: The textbook*, pp. 1–28. Springer, 2016.

Majid Hameed Ahmed, Sabrina Tiun, Nazlia Omar, and Nor Samsiah Sani. Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*, 13(1):342, 2022.

Koenraad M R Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127, jun 2007. doi: 10.1088/1751-8113/40/28/S18.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1998–2013, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.154. URL https://aclanthology.org/2022.findings-naacl.154/.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pp. 594–602. SIAM, 2019.

Qiang Ding, Yixuan Cao, and Ping Luo. Top-ambiguity samples matter: understanding why deep ensemble works in selective classification. *Advances in Neural Information Processing Systems*, 36:35497–35521, 2023.

Jingcan Duan, Siwei Wang, Pei Zhang, En Zhu, Jingtao Hu, Hu Jin, Yue Liu, and Zhibin Dong. Graph anomaly detection via multi-scale contrastive learning networks with augmented view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7459–7467, 2023.

Meysam Ghaffari and Nasser Ghadiri. Ambiguity-driven fuzzy c-means clustering: how to detect uncertain clustered records. *Applied Intelligence*, 45(2):293–304, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. Embed2detect: temporally clustered embedded words for event detection in social media. *Machine Learning*, 111(1):49–87, 2022.

Anis Hoayek and Didier Rullière. Assessing clustering methods using shannon's entropy. *Information Sciences*, 689:121510, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2282952.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. Anemone: Graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 3122–3126, 2021.

Krissanee Kamthawee, Can Udomcharoenchaikit, and Sarana Nutanong. Mist: mutual information maximization for short text clustering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11309–11324, 2024.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1), February 2015. ISSN 0090-5364. doi: 10.1214/14-aos1274.

Jia Li, Jiangwei Li, Chenxu Wang, Fons J Verbeek, Tanja Schultz, and Hui Liu. Ms2od: outlier detection using minimum spanning tree and medoid selection. *Machine Learning: Science and Technology*, 5(1):015025, 2024a.

Zetong Li, Qinliang Su, Shijing Si, and Jianxing Yu. Leveraging bert and tfidf features for short text clustering via alignment-promoting co-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14897–14913, 2024b.

Zetong Li, Qinliang Su, Minhua Huang, and Yin Yang. Co-evolving llms and embedding models via density-guided preference optimization for text clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4796–4808, 2025.

I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. Twist: Training-free and label-free short text clustering through iterative vector updating with llms. *arXiv preprint arXiv:2510.06747*, 2025.

Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2378–2392, June 2022. ISSN 2162-2388. doi: 10.1109/tnnls.2021.3068344.

Yixin Liu, Kaize Ding, Huan Liu, and Shirui Pan. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 339–347, 2023.

Yonghao Liu, Fausto Giunchiglia, Lan Huang, Ximing Li, Xiaoyue Feng, and Renchu Guan. A simple graph contrastive learning framework for short text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19015–19023, 2025.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Junjun Pan, Yixin Liu, Yizhen Zheng, and Shirui Pan. Prem: A simple yet effective approach for node-level graph anomaly detection. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1253–1258. IEEE, 2023.

Huan Qing. Overlapping community detection in weighted networks. *Statistical Papers*, 66(6), October 2025. ISSN 1613-9798. doi: 10.1007/s00362-025-01758-y.

Huan Qing and Jingli Wang. Bipartite mixed membership distribution-free model. a novel model for community detection in overlapping bipartite weighted networks, 2024a. URL https://arxiv.org/abs/2211.00912.

Huan Qing and Jingli Wang. Mixed membership distribution-free model, 2024b. URL https://arxiv.org/abs/2112.04389.

Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. Enhancement of short text clustering by iterative classification, 2020. URL https://arxiv.org/abs/2001.11631.

Geoffrey Schiebinger, Martin J. Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2), April 2015. ISSN 0090-5364. doi: 10.1214/14-aos1283.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, et al. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web*, 19(3):1–29, 2025.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Yu Tian, Zachary Lubberts, and Melanie Weber. Mixed-membership community detection via line graph curvature. In Sophia Sanborn, Christian Shewmake, Simone Azeglio, Arianna Di Bernardo, and Nina Miolane (eds.), *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197 of *Proceedings of Machine Learning Research*, pp. 219–233. PMLR, 03 Dec 2023.

Jamalbek Tussupov, Akmaral Kassymova, Ayagoz Mukhanova, Assyl Bissengaliyeva, Zhanar Azhibekova, Moldir Yessenova, and Zhanargul Abuova. Analysis of short texts using intelligent clustering methods. *Algorithms*, 18(5):289, 2025.

Ulrike von Luxburg. A tutorial on spectral clustering, 2007.

Ulrike Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.

Tu Vu, Manh Do, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. Topic modeling for short texts via optimal transport-based clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7666–7680, 2025.

Hongtao Wang, Taiyan Zhang, Renchi Yang, and Jianliang Xu. Cequel: Cost-effective querying of large language models for text clustering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 2998–3008, 2025a.

Ying Wang, Mengye Ren, and Andrew Gordon Wilson. In-context clustering with large language models. *arXiv preprint arXiv:2510.08466*, 2025b.

Hui Wu, Yuanben Zhang, Zhonghe Han, Yingyan Hou, Lei Wang, Siye Liu, Qihang Gong, and Yunping Ge. Cot-driven framework for short text classification: Enhancing and transferring capabilities from large to smaller model. *Knowledge-Based Systems*, 311:113057, 2025.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*, 2023.

Chen Yang, Bin Cao, and Jing Fan. Tec: A novel method for text clustering with large language models guidance and weakly-supervised contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 1702–1712, 2024.

Zhihao Yao. Discriminative representation learning via attention-enhanced contrastive learning for short text clustering. *arXiv preprint arXiv:2501.03584*, 2025.

Zhihao Yao, Jixuan Yin, and Bo Li. Reliable pseudo-labeling via optimal transport with attention for short text clustering. *arXiv preprint arXiv:2501.15194*, 2025.

Jiaqiang Zhang, Senzhang Wang, and Songcan Chen. Reconstruction enhanced multi-view contrastive learning for anomaly detection on attributed networks. *arXiv preprint arXiv:2205.04816*, 2022.

Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12220–12233, 2021.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. Simans: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 548–559, 2022.

Ruoyu Zhou, Lianjie Shu, and Yan Su. An adaptive minimum spanning tree test for detecting irregularly-shaped spatial clusters. *Computational Statistics & Data Analysis*, 89:134–146, 2015.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Computing Surveys*, 57(3):1–38, 2024.

Henry Peng Zou, Siffi Singh, Yi Nian, Jianfeng He, Jason Cai, Saab Mansour, and Hang Su. Glean: Generalized category discovery with diverse and quality-enhanced llm feedback. *arXiv preprint arXiv:2502.18414*, 2025.

# A NOTATIONS

Appendix A provides a table summarizing the key notation used throughout the paper, while Appendix B presents two supporting lemmas and the proofs of the theoretical results.

Table 4: Notations

| Notation | Description |
|---|---|
| $n$ | total number of texts |
| $\mathcal{V} = \{1, \ldots, n\}$ | entire sample set |
| $d$ | embedding dimension |
| $x_i \in \mathbb{R}^d$ | embedding vector of sample $i$ |
| $\delta_{ij}$ | cosine distance between embeddings $x_i$ and $x_j$ |
| $w_{ij}$ | semantic similarity weight associated with edge $(i, j)$ |
| $K$ | number of nearest neighbors in the kNN graph |
| $\mathcal{N}_K(i)$ | $K$-nearest-neighbor set of sample $i$ |
| $\mathcal{E}_K$ | edge set of the kNN graph |
| $G_K = (\mathcal{V}, \mathcal{E}_K)$ | weighted kNN semantic graph |
| $T = \mathrm{MST}(G_K)$ | minimum spanning tree of $G_K$ |
| $\Phi(H)$ | edge-removal fragility of a connected graph $H$ |
| $d_{ij}^{\mathrm{MST}}$ | distance of edge $(i, j)$ on the MST |
| $p_{ij}$ | probability of cutting MST edge $(i, j)$ |
| $P$ | distance-aware probability distribution over MST edges |
| $k$ | number of semantic clusters |
| $\mathcal{C}(T) = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ | reference partition obtained by removing the $k-1$ longest edges of $T$ |
| $\mathcal{C}^{(\ell)}(T)$ | partition of $T$ obtained at perturbation round $\ell$ |
| $S^{(\ell)} \subseteq \mathcal{V}$ | sampled connected subgraph at round $\ell$ |
| $L$ | total number of perturbation rounds |
| $L_i$ | number of rounds in which sample $i$ is sampled |
| $r^{(\ell)}$ | number of spectral eigenvectors used at round $\ell$ |
| $z^{(\ell)}$ | raw clustering labels on $S^{(\ell)}$ |
| $\hat{z}^{(\ell)}$ | clustering labels after Hungarian alignment |
| $\hat{\pi}_{ic}$ | empirical assignment frequency of sample $i$ to label $c$ |
| $\hat{\pi}_i \in \Delta^{k-1}$ | empirical assignment distribution of sample $i$ |
| $\pi_{ic}$ | conditional population assignment probability to label $c$ |
| $\pi_i \in \Delta^{k-1}$ | conditional population assignment distribution of sample $i$ |
| $\hat{H}_i$ | normalized empirical assignment entropy of sample $i$ |
| $\tau$ | entropy threshold separating intra-cluster and ambiguous samples |
| $h_{\mathrm{in}}$ | asymptotic upper entropy level for intra-cluster samples |
| $h_{\mathrm{amb}}$ | asymptotic lower entropy level for ambiguous samples |
| $\mathcal{T} = \bigcup_{r=1}^{k} \mathcal{T}_r$ | set of intra-cluster samples |
| $\mathcal{T}_r$ | tight semantic cluster $r$ |
| $\mathcal{F}$ | set of ambiguous samples |
| $a(i), b(i)$ | two nearby semantic clusters associated with ambiguous sample $i$ |
| $\mathcal{T}_{r(i)}$ | nearby cluster of ambiguous sample $i$, $r \in \{a, b\}$ |
| $\Delta_{\mathrm{in}}$ | upper bound on intra-cluster distances |
| $\Delta_{\mathrm{out}}$ | lower bound on inter-cluster distances |
| $\Delta_{\mathrm{ambiguous}}$ | minimum separation of ambiguous samples from cluster interiors |
| $\Delta_{\mathrm{near}}$ | proximity threshold to nearby semantic clusters |
| $B_i^r \subseteq E(T)$ | boundary-switching MST edge set of sample $i$ toward $\mathcal{T}_{r(i)}$ |
| $\Phi_i^r(H)$ | local perturbation effectiveness toward $\mathcal{T}_{r(i)}$ in graph $H$ |
| $C^{(\ell)}$ | set of MST edges removed at perturbation round $\ell$ |
| $E_i^{r,(\ell)}$ | exclusive activation event of $B_i^r$ at round $\ell$ |
| $q_i^{\mathrm{FRGP}}$ | minimum effective boundary-switching probability of sample $i$ under FRGP |
| $q_*^{\mathrm{FRGP}}$ | asymptotic uniform lower bound of $q_i^{\mathrm{FRGP}}$ over ambiguous samples |
| $p_{\min}$ | minimum per-round sampling probability of any sample |
| $\mathcal{A}^{(\ell)}$ | successful clustering and label-alignment event at round $\ell$ |
| $\varepsilon_n$ | conditional clustering/alignment error bound, $\varepsilon_n \to 0$ |
| $I_i^{(\ell)}$ | indicator that sample $i$ is included at round $\ell$ |
| $Z_i^{(\ell)}$ | aligned label of sample $i$ at round $\ell$ when sampled |
| $\beta_{i,\ell}^{r,M}$ | conditional exclusive boundary-event probability under mechanism $M$ |
| $q_i^{(M)}$ | effective boundary-switching probability of sample $i$ under mechanism $M$ |
| $q_{\min,n}^{(M)}$ | minimum of $q_i^{(M)}$ over ambiguous samples |
| $\Gamma_M$ | theoretical entropy-separation lower bound for mechanism $M$ |

Continued on next page

Table 4 – continued

| Notation | Description |
|---|---|
| $r_i(\delta)$ | finite-sample $\ell_1$ concentration radius for $\hat{\pi}_i$ |
| $g(\cdot)$ | normalized entropy-continuity bound from Fannes–Audenaert |
| $h_2(\cdot)$ | binary entropy function |
| $\bar{s}, \bar{m}$ | average sampled-subgraph size and edge count |
| $t, I$ | eigensolver and $k$-means iteration counts |

# B Supporting Lemmas and Proofs of the Theoretical Results

**Boundary events and notation.** For a graph $H$ on $\mathcal{V}$, write $\mathcal{R}_i^r(H) = \{v \in \mathcal{T}_{r(i)} : i \leftrightarrow v \text{ in } H\}$, where $r \in \{a, b\}$ indexes a neighboring label. Put

$$D_i^r(H, e) = \mathbf{1}\{\mathcal{R}_i^r(H - e) \neq \mathcal{R}_i^r(H)\}, \qquad \Phi_i^r(H) = \sum_{e \in B_i^r} \nu_i^r(e) D_i^r(H, e), \tag{6}$$

where $\nu_i^r$ is a specified probability law on the *same* set $B_i^r \subseteq E(T) \subseteq E(G_K)$ for both graphs. Thus local fragility compares matched boundary proposals, not independently normalized uniform draws from different edge sets. Also $\Phi(H) = |\operatorname{Br}(H)|/|E(H)|$, with $\operatorname{Br}(H)$ the bridges of a connected graph with at least one edge.

Let $C^{(\ell)}$ be the full set of edges removed in a FRGP round and $Z_i^{(\ell)} \in [k]$ the aligned label when $I_i^{(\ell)} := \mathbf{1}\{i \in S^{(\ell)}\} = 1$. For $t \neq r$ in $\{a, b\}$, define the exclusive boundary activation

$$E_i^{r,(\ell)} = \{C^{(\ell)} \cap B_i^r \neq \varnothing,\ C^{(\ell)} \cap B_i^t = \varnothing\}, \quad \beta_{i,\ell}^{r,\mathrm{FRGP}} = \mathbb{P}(E_i^{r,(\ell)} \mid I_i^{(\ell)} = 1). \tag{7}$$

This formalizes *activating* $B_i^r$ *alone* in Assumption 4.3. The two events are disjoint even if the boundary sets overlap. They do not include the success event $A^{(\ell)}$. Write

$$\beta_i^{r,\mathrm{FRGP}} = \min_{\ell \in [L]} \beta_{i,\ell}^{r,\mathrm{FRGP}}, \qquad q_i^{\mathrm{FRGP}} = \min_{r \in \{a,b\}} \beta_i^{r,\mathrm{FRGP}} - \varepsilon_n. \tag{8}$$

Neither positive coverage nor a large unconditional hit probability guarantees positive $q_i^{\mathrm{FRGP}}$.

## B.1 Useful Lemmas

**Lemma B.1** (Conditional concentration of empirical votes)**.** *Under the independent-round implementation, let* $\mathbf{I}_i = (I_i^{(1)}, \ldots, I_i^{(L)})$ *and define* $\pi_i$ *by equation 11. For any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, every node with* $L_i > 0$ *satisfies*

$$\|\widehat{\pi}_i - \pi_i\|_1 \leq k \sqrt{\frac{\log(2nk/\delta)}{2L_i}}.$$

*The rounds need not be identically distributed.*

*Proof.* Fix $i$ and condition on any positive-probability value of $\mathbf{I}_i$. Round-wise independence implies that conditioning on the separate events $I_i^{(\ell)} = 0$ or 1 preserves independence of the selected round outputs. For each $c$, the selected variables $Y_\ell = \mathbf{1}\{Z_i^{(\ell)} = c\}$ are independent and lie in $[0, 1]$, with means $p_{ic}^{(\ell)} = \mathbb{P}(Z_i^{(\ell)} = c \mid I_i^{(\ell)} = 1)$. Consequently, if $L_i > 0$, Hoeffding's inequality (Hoeffding, 1963) gives

$$\mathbb{P}(|\widehat{\pi}_{ic} - \pi_{ic}| > u \mid \mathbf{I}_i) \leq 2\exp(-2L_i u^2).$$

For $u_i = \sqrt{\log(2nk/\delta)/(2L_i)}$ the right-hand side is $\delta/(nk)$. Sum over $c$ to get conditional failure probability at most $\delta/n$ for $\|\widehat{\pi}_i - \pi_i\|_1 \leq k u_i$. Average over $\mathbf{I}_i$, counting failure only on $L_i > 0$, to obtain the same unconditional bound $\delta/n$. Finally sum over $i$. No independence across nodes is needed. In particular, we do not condition jointly on all nodes' inclusion vectors, which could alter the individual conditional means. □

**Lemma B.2** (Entropy continuity)**.** *For probability vectors $p, q$ on $k \geq 2$ labels, write $H(p) = -\sum_c p_c \log p_c / \log k$ and $d = \|p - q\|_1/2$. Then*

$$|H(p) - H(q)| \leq \frac{h_2(d) + d \log(k-1)}{\log k}.$$

*In particular, if $\|p - q\|_1 \leq t \leq 2(1 - 1/k)$, the right-hand side is at most $g(t)$.*

*Proof.* Apply the Fannes–Audenaert inequality (Audenaert, 2007) to the diagonal density matrices with diagonals $p$ and $q$. Their trace distance is $d$ and their von Neumann entropies equal the corresponding unnormalized Shannon entropies. Division by $\log k$ proves the first claim. To replace the exact distance by an upper bound, set $f(d) = h_2(d) + d \log(k-1)$. For $0 < d < 1$,

$$f'(d) = \log \frac{(k-1)(1-d)}{d} \geq 0 \quad \Longleftrightarrow \quad d \leq 1 - 1/k.$$

Thus $d \leq t/2 \leq 1 - 1/k$ implies $f(d) \leq f(t/2)$. The endpoint cases follow by continuity. The monotonicity restriction is needed only when inserting a distance upper bound. □

### B.2 Details of Definition 3.1

*Proof.* Recall that $\Phi(H) = |\{e \in \mathcal{E}_H : e \text{ is a bridge}\}|/|\mathcal{E}_H|$.

($\Rightarrow$) If $H$ is a spanning tree of $G_K$, then it is connected and acyclic. In a tree, every edge is a bridge, since removing any edge disconnects it into two components. Hence $\Phi(H) = 1$.

($\Leftarrow$) Suppose $\Phi(H) = 1$, i.e., every edge of $H$ is a bridge. If $H$ contained a cycle, then no edge on that cycle would be a bridge, contradicting $\Phi(H) = 1$. Therefore $H$ is acyclic. Since $H$ is a spanning subgraph of the connected graph $G_K$ and has no cycles, it is a spanning tree.

Since all spanning trees of $G_K$ satisfy $\Phi(H) = 1$, we select $T^* = \arg\min_{T \in \mathcal{T}(G_K)} \sum_{(i,j) \in T} \delta_{ij}$ to additionally minimize the total semantic distance. □

### B.3 Proof of Proposition 4.6

**Proposition B.3** (MST maximizes perturbation effectiveness)**.** *For any connected graph $H$, uniformly deleting one edge disconnects $H$ with probability $\Phi(H)$. For $T = \mathrm{MST}(G_K)$, $\Phi(T) = 1$, whereas $\Phi(G_K) < 1$ whenever $G_K$ contains a cycle. For the matched local law in equation 6, if $\nu_i^r(\{e : D_i^r(G_K, e) = 0\}) > 0$, then $\Phi_i^r(T) = 1 > \Phi_i^r(G_K)$. This holds for each $r \in \{a, b\}$ satisfying the condition.*

*Proof.* For a connected graph $H$, an edge is a bridge precisely when deleting it disconnects $H$. Under uniform edge selection,

$$\mathbb{P}(H - e \text{ is disconnected}) = \sum_{e \in E(H)} \frac{\mathbf{1}\{e \in \mathrm{Br}(H)\}}{|E(H)|} = \Phi(H).$$

Every edge in a tree is a bridge, and hence

$$\Phi(T) = 1.$$

If $G_K$ contains a cycle, every edge on that cycle is non-bridging, so

$$|\mathrm{Br}(G_K)| < |E(G_K)|$$

and therefore

$$\Phi(G_K) < 1.$$

For the local statement, Assumption 4.3 and the reachability definition give

$$D_i^r(T, e) = 1, \qquad e \in B_i^r.$$

Since $\sum_e \nu_i^r(e) = 1$,

$$\Phi_i^r(T) = 1, \qquad \Phi_i^r(G_K) = 1 - \nu_i^r(\{e \in B_i^r : D_i^r(G_K, e) = 0\}).$$

Thus, if

$$\nu_i^r(\{e \in B_i^r : D_i^r(G_K, e) = 0\}) > 0,$$

then

$$\Phi_i^r(T) > \Phi_i^r(G_K).$$

Here $D_i^r(G_K, e) = 0$ means that removing $e$ leaves the reachable set $\mathcal{R}_i^r(G_K)$ unchanged. □

### B.4 Proof of Proposition 4.7

**Proposition B.4** (Distance-aware cutting favors ambiguous boundary edges)**.** *For a single draw on $T$, let $m = |E(T)|$, $w_e = d_e^{\mathrm{MST}}$, $W = \sum_{e \in E(T)} w_e > 0$, and $P_{\mathrm{FRGP}}(e) = w_e/W$. For $B = B_i^r$, put $b = |B|$ and $W_B = \sum_{e \in B} w_e$. Then*

$$\frac{W_B}{b} > \frac{W}{m} \quad \Longrightarrow \quad P_{\mathrm{FRGP}}(B) = \frac{W_B}{W} > \frac{b}{m} = P_{\mathrm{U}}(B). \tag{9}$$

*For $s = k - 1$ sequential draws without replacement, the probability of at least one hit on $B$ is also strictly larger under FRGP if $1 \le s \le m - b$ and, on every positive-probability FRGP history of $j < s$ misses,*

$$\frac{W_B}{W - \sum_{e \in C_j} w_e} \ge \frac{b}{m - j}, \tag{10}$$

*where $C_j \subseteq E(T) \setminus B$ is the set already removed. The average-length condition alone establishes the single-draw comparison; the sequential result additionally requires equation 10.*

For $L_i = \sum_\ell I_i^{(\ell)} > 0$, define

$$\widehat{\pi}_{ic} = \frac{1}{L_i} \sum_{\ell : I_i^{(\ell)} = 1} \mathbf{1}\{Z_i^{(\ell)} = c\}, \qquad \pi_{ic} = \frac{1}{L_i} \sum_{\ell : I_i^{(\ell)} = 1} \mathbb{P}(Z_i^{(\ell)} = c \mid I_i^{(\ell)} = 1), \tag{11}$$

$$\widehat{H}_i = \frac{-\sum_{c=1}^k \widehat{\pi}_{ic} \log \widehat{\pi}_{ic}}{\log k}.$$

We use natural logarithms and $0 \log 0 = 0$.

*Proof.* For one draw,

$$P_{\mathrm{FRGP}}(B) - P_{\mathrm{U}}(B) = \frac{W_B}{W} - \frac{b}{m} = \frac{b}{W}\left(\frac{W_B}{b} - \frac{W}{m}\right) > 0,$$

which is exactly the average-edge-length condition in Assumption 4.3.

For sequential draws, let $N_j$ be the event of no hit on $B$ in the first $j$ cuts. Given a positive-probability FRGP history $C_j$ on $N_j$, the next-cut hit probability is

$$\frac{W_B}{W - \sum_{e \in C_j} w_e},$$

whereas under uniform cutting it is

$$\frac{b}{m - j}.$$

Under equation 10,

$$\mathbb{P}_{\mathrm{FRGP}}(N_{j+1} \mid N_j) \le 1 - \frac{b}{m - j}.$$

Multiplying the conditional survival probabilities and using the strict first-cut inequality gives

$$\mathbb{P}_{\mathrm{FRGP}}(N_s) \le \left(1 - \frac{W_B}{W}\right) \prod_{j=1}^{s-1} \left(1 - \frac{b}{m - j}\right) < \prod_{j=0}^{s-1} \left(1 - \frac{b}{m - j}\right) = \mathbb{P}_{\mathrm{U}}(N_s).$$

Subtracting from one yields

$$P_{\mathrm{FRGP}}(B \text{ is hit}) > P_{\mathrm{U}}(B \text{ is hit}),$$

which proves the sequential claim. □

### B.5 Proof of Theorem 4.8

*Proof.* We first bound the conditional population entropy and then control its empirical deviation.

Fix $i \in \mathcal{T}_r$ and a round $\ell$. On $I_i^{(\ell)} = 1$, Assumption 4.5 implies

$$\mathbb{P}(Z_i^{(\ell)} \neq r \mid I_i^{(\ell)} = 1) \leq \mathbb{P}((A^{(\ell)})^c \mid I_i^{(\ell)} = 1) \leq \varepsilon_n.$$

Averaging over the selected rounds gives $\eta_i := 1 - \pi_{ir} \leq \varepsilon_n$. If $\eta_i > 0$, let $v_c = \pi_{ic}/\eta_i$ for $c \neq r$. Then

$$-\sum_c \pi_{ic} \log \pi_{ic} = h_2(\eta_i) + \eta_i \left( -\sum_{c \neq r} v_c \log v_c \right) \leq h_2(\eta_i) + \eta_i \log(k-1).$$

The case $\eta_i = 0$ follows by continuity. For sufficiently large $n$, $\varepsilon_n \leq 1 - 1/k$, and the monotonicity established in Lemma B.2 yields

$$H(\pi_i) \leq \frac{h_2(\varepsilon_n) + \varepsilon_n \log(k-1)}{\log k} = g(2\varepsilon_n). \tag{12}$$

Now fix $i \in \mathcal{F}$, $r \in \{a, b\}$, and $\ell$. By Assumption 4.3, on $I_i^{(\ell)} = 1$ the event $E_i^{r,(\ell)} \cap A^{(\ell)}$ implies $Z_i^{(\ell)} = r(i)$. Hence

$$\begin{aligned} \mathbb{P}(Z_i^{(\ell)} = r(i) \mid I_i^{(\ell)} = 1) &\geq \mathbb{P}(E_i^{r,(\ell)} \cap A^{(\ell)} \mid I_i^{(\ell)} = 1) \\ &\geq \beta_{i,\ell}^{r,\mathrm{FRGP}} - \varepsilon_n \\ &\geq q_i^{\mathrm{FRGP}}. \end{aligned}$$

Averaging over the selected rounds gives

$$\pi_{i,a(i)}, \pi_{i,b(i)} \geq q_i^{\mathrm{FRGP}}.$$

Since the exclusive events for $a$ and $b$ are disjoint, $q_i^{\mathrm{FRGP}} \leq 1/2$.

Let $q = q_i^{\mathrm{FRGP}} > 0$ and $x = \pi_{i,a(i)}$. Since $\pi_{i,b(i)} \geq q$, we have $q \leq x \leq 1 - q$. Grouping all labels other than $a(i)$ and denoting their normalized distribution by $v$, we obtain

$$-\sum_c \pi_{ic} \log \pi_{ic} = h_2(x) + (1-x) \left( -\sum_c v_c \log v_c \right) \geq h_2(x) \geq h_2(q).$$

Therefore

$$H(\pi_i) \geq \frac{h_2(q_i^{\mathrm{FRGP}})}{\log k}. \tag{13}$$

By Lemma B.1, with probability at least $1 - \delta$, simultaneously for all $i$ with $L_i > 0$,

$$\|\widehat{\pi}_i - \pi_i\|_1 \leq r_i(\delta).$$

Whenever $r_i(\delta) \leq 2(1 - 1/k)$, Lemma B.2 gives

$$|\widehat{H}_i - H(\pi_i)| \leq g(r_i(\delta)).$$

Combining this inequality with equation 12 and equation 13 proves equation 3–equation 4.

It remains to establish the asymptotic separation. Choose $q_0$ such that $0 < q_0 < q_*^{\mathrm{FRGP}} \leq 1/2$. For all sufficiently large $n$,

$$\min_{i \in \mathcal{F}} q_i^{\mathrm{FRGP}} \geq q_0.$$

Let

$$c = \frac{h_2(q_0)}{\log k} > 0.$$

Since $\varepsilon_n \to 0$, $g(2\varepsilon_n) \to 0$. Moreover, for any fixed $\delta > 0$, the count condition and fixed $k$ imply

$$R_n(\delta) := \max_i r_i(\delta) = k \sqrt{\frac{\log(2nk/\delta)}{2 \min_i L_i}} \xrightarrow{p} 0.$$

Hence $g(R_n(\delta)) \xrightarrow{p} 0$. Set, for example,

$$h_{\text{in}} = \frac{c}{4}, \qquad h_{\text{amb}} = \frac{3c}{4}, \qquad \tau = \frac{c}{2}.$$

Then, with probability tending to one,

$$\max_{i \in \mathcal{T}} \widehat{H}_i \le h_{\text{in}} < \tau < h_{\text{amb}} \le \min_{i \in \mathcal{F}} \widehat{H}_i.$$

Therefore

$$\mathbb{P}\Big( \max_{i \in \mathcal{T}} \widehat{H}_i \le \tau,\ \min_{i \in \mathcal{F}} \widehat{H}_i \ge \tau \Big) \longrightarrow 1,$$

which proves equation 2. □

### B.6 Proof of Corollary 4.9

For each mechanism $M$, let $I_i^{(\ell,M)}$ denote the inclusion indicator and $E_i^{r,(\ell,M)}$ an exclusive boundary event such that, on successful clustering and alignment, it induces label $r(i)$. Define

$$\beta_{i,\ell}^{r,M} = \mathbb{P}_M\Big( E_i^{r,(\ell,M)} \mid I_i^{(\ell,M)} = 1 \Big).$$

The comparison assumes a common conditional clustering/alignment error bound $\varepsilon_n$ across mechanisms. For each mechanism $M$, define

$$q_i^{(M)} = \min_{r \in \{a,b\},\, \ell \in [L]} \beta_{i,\ell}^{r,M} - \varepsilon_n.$$

These conditions formalize the matched-ablation design used in our experiments: A1 removes the MST backbone while keeping the downstream clustering/alignment procedure and round budget unchanged, whereas A2 keeps the same MST and replaces distance-aware cutting only with uniform cutting. Thus, each ablation isolates one component of FRGP.

**A1: matched local proposals.** For every $i, r, \ell$, suppose FRGP and A1 use the same local boundary proposal law $\nu_i^r$ and the same positive downstream factor $a_{i,\ell}^r$, so that

$$\beta_{i,\ell}^{r,\text{FRGP}} = a_{i,\ell}^r \Phi_i^r(T), \qquad \beta_{i,\ell}^{r,\text{A1}} = a_{i,\ell}^r \Phi_i^r(G_K). \tag{14}$$

Hence the comparison isolates the effect of replacing the MST by the original kNN graph, and Proposition 4.6 transfers the local fragility advantage to the effective boundary probability.

**A2: matched hit-to-vote conversion.** Let $J_i^{r,(\ell,M)}$ denote the event that at least one cut hits $B_i^r$, and let

$$h_{i,\ell}^{r,M} = \mathbb{P}_M\Big( J_i^{r,(\ell,M)} \Big).$$

Under the conditions of Proposition 4.7, suppose the conversion from a boundary hit to an exclusive sampled boundary event is shared by FRGP and A2:

$$b_{i,\ell}^r = \frac{\mathbb{P}_M\Big( E_i^{r,(\ell,M)} \cap \{ I_i^{(\ell,M)} = 1 \} \mid J_i^{r,(\ell,M)} \Big)}{\mathbb{P}_M\Big( I_i^{(\ell,M)} = 1 \Big)} > 0, \qquad M \in \{\text{FRGP}, \text{A2}\}. \tag{15}$$

Then

$$\beta_{i,\ell}^{r,M} = b_{i,\ell}^r h_{i,\ell}^{r,M}. \tag{16}$$

Thus A2 isolates the effect of distance-aware versus uniform MST cutting, while the remaining conversion from a boundary hit to an effective vote is matched.

Under these matched-ablation conditions, Propositions 4.6 and 4.7 imply the conditional probability comparisons used in Corollary 4.9.

*Proof.* Fix $n$ and the finite round count $L$. Under the local bypass condition, Proposition 4.6 and equation 14 imply, for every $i, r, \ell$,

$$\beta_{i,\ell}^{r,\mathrm{FRGP}} - \beta_{i,\ell}^{r,\mathrm{A1}} = a_{i,\ell}^{r} \left\{ \Phi_i^r(T) - \Phi_i^r(G_K) \right\} > 0.$$

Likewise, Proposition 4.7 and equation 16 give

$$\beta_{i,\ell}^{r,\mathrm{FRGP}} - \beta_{i,\ell}^{r,\mathrm{A2}} = b_{i,\ell}^{r} \left\{ h_{i,\ell}^{r,\mathrm{FRGP}} - h_{i,\ell}^{r,\mathrm{A2}} \right\} > 0.$$

For $M = \mathrm{A1}$ or A2, finiteness and strictness at every index yield

$$d_{n,M} := \min_{i \in \mathcal{F},\ r \in \{a,b\},\ \ell \in [L]} \left( \beta_{i,\ell}^{r,\mathrm{FRGP}} - \beta_{i,\ell}^{r,M} \right) > 0.$$

Consequently,

$$q_i^{\mathrm{FRGP}} \geq q_i^{(M)} + d_{n,M}, \qquad q_{\min,n}^{\mathrm{FRGP}} \geq q_{\min,n}^{(M)} + d_{n,M} > q_{\min,n}^{(M)}.$$

As in Appendix B.5, the two exclusive boundary events yield two label masses of at least $q_i^{(M)}$, and therefore

$$H(\pi_i) \geq \frac{h_2(q_i^{(M)})}{\log k}, \qquad i \in \mathcal{F}.$$

For intra-cluster samples, the common clustering/alignment error bound gives

$$H(\pi_i) \leq g(2\varepsilon_n), \qquad i \in \mathcal{T}.$$

Hence

$$\Gamma_M = \frac{h_2(q_{\min,n}^{(M)})}{\log k} - g(2\varepsilon_n)$$

is a lower bound on the population entropy separation.

Since $h_2$ is strictly increasing on $[0, 1/2]$,

$$\Gamma_{\mathrm{FRGP}} - \Gamma_M = \frac{h_2(q_{\min,n}^{\mathrm{FRGP}}) - h_2(q_{\min,n}^{(M)})}{\log k} > 0.$$

Applying this to $M = \mathrm{A1}$ and $M = \mathrm{A2}$ proves equation 5. □

# C Details of simulation studies

## C.1 Simulation settings

We conduct simulations to evaluate FRGP in a controllable setting with ground-truth ambiguity labels. This allows us to examine the theoretically established entropy-based separability of FRGP in Theorem 4.8. Simulation datasets are generated in a $d$-dimensional embedding space with $k$ semantic clusters; unless otherwise specified, we use $d = 128$, $k = 5$ tight clusters, and $n = 20000$.

**Generation of intra-cluster samples.** Intra-cluster samples are generated from $k$ Gaussian clusters that represent stable semantic cores. The cluster centers $\{\mu_c\}_{c=1}^{k}$ are separated by at least $\Delta$, with the default setting $\Delta = 50$ to ensure clear inter-cluster separation. Each intra-cluster sample follows $x \sim \mathcal{N}(\mu_c, \Sigma_c)$, where the intra-cluster variance is small, $\sigma_{\text{tight}}^2 = 0.1^2$. To avoid overly idealized spherical clusters, we apply dimension-wise scaling to $\Sigma_c$, introducing anisotropic covariance structures and more realistic local boundary geometry.

**Generation of ambiguous samples.** For two randomly selected tight clusters $a$ and $b$, we draw $\mathbf{x}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ and $\mathbf{x}_b \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$, and generate an ambiguous sample as $\mathbf{x} = \lambda \mathbf{x}_a + (1-\lambda)\mathbf{x}_b$, where $\lambda \in [0, 1]$. This captures mixed-membership ambiguous samples.

**Experimental evaluation.** To evaluate FRGP systematically, we vary four key factors: (i) sample size $n$, (ii) ambiguous sample ratio $r_{\text{amb}}$, (iii) ambiguity strength $\lambda$, and (iv) separability of the cluster $\Delta$. In each sweep, we vary one factor and keep others fixed. For each configuration, FRGP computes node-wise entropy scores $\hat{H}_i$ on the generated embeddings, and the top-ranked samples matching the ground-truth ambiguous ratio are selected as predicted ambiguous samples. Each configuration is repeated five times with independent random seeds, and mean results are reported. The ground-truth proportion defines an evaluation operating point only; it is not an input to FRGP at deployment.

### C.2 SIMULATION RESULTS

**Effect of sample size.** Table 5 in Appendix C reports performance under increasing sample sizes. Both AUC and Precision improve monotonically as $n$ increases, approaching accurate identification. Larger networks expose ambiguous samples to more diverse local neighborhoods, causing more frequent assignment fluctuations and clearer entropy separation from tight clusters.

**Effect of ambiguous sample ratio.** As shown in Table 5, increasing the proportion of ambiguous samples leads to substantial gains in identification performance. When ambiguous samples are rare, their instability signals may be diluted by dominant tight cluster structures.

**Effect of ambiguity strength.** Table 5 shows that performance peaks when $\lambda \approx 0.5$, where ambiguous samples lie near inter-cluster midpoints, and decreases as $\lambda$ approaches $0$ or $1$, where samples move closer to a single cluster core. This non-monotonic pattern supports the intuition that semantic ambiguity is strongest near cluster boundaries.

**Effect of cluster separability.** Table 5 shows that FRGP remains effective across separability levels. As $\Delta$ increases, tight clusters become more stable, while ambiguous samples remain perturbation-sensitive near inter-cluster bottlenecks, indicating that FRGP captures local structural fragility rather than relying on extreme cluster overlap.

Table 5: Simulation results for ambiguous samples identification under different parameter sweeps.

| Sweep type | Setting | AUC | Precision |
|---|---|---|---|
| Sample size $n$ | 1000 | 0.7483 | 0.2267 |
| | 2000 | 0.8602 | 0.2967 |
| | 5000 | 0.9513 | 0.6307 |
| | 10000 | 0.9891 | 0.9420 |
| | 20000 | **1.0000** | **1.0000** |
| Ambiguous ratio $r_{\text{amb}}$ | 0.01 | 0.6538 | 0.0467 |
| | 0.02 | 0.7533 | 0.0633 |
| | 0.05 | 0.9537 | 0.6813 |
| | 0.10 | 0.9995 | 0.9667 |
| | 0.20 | **1.0000** | **1.0000** |
| Ambiguity $\lambda$ | 0.00 | 0.9586 | 0.7590 |
| | 0.05 | 0.9515 | 0.7620 |
| | 0.10 | 0.9198 | 0.6677 |
| | 0.30 | 0.9190 | 0.7280 |
| | 0.50 | **1.0000** | **1.0000** |
| | 0.70 | 0.9227 | 0.7157 |
| | 0.90 | 0.8714 | 0.6602 |
| | 0.95 | 0.9694 | 0.7930 |
| | 1.00 | 0.9788 | 0.8043 |
| Separability $\Delta$ | 30 | 0.8212 | 0.3307 |
| | 70 | 0.9359 | 0.6000 |
| | 90 | 0.9670 | 0.6213 |
| | 110 | **0.9782** | **0.6240** |

## D ADDITIONAL DETAILS OF EXPERIMENTS

**Semantic-description prompt.** For each dish name, we query DeepSeek V3.2 using the following prompt, where "example dish name" is replaced with the input dish name: "Please analyze 'example dish name' in no more than about 20 words, and determine which category of food or beverage it belongs to, or whether it is some kind of descriptive instruction." This prompt elicits a concise category-level description while allowing the model to identify descriptive or instructional expressions rather than forcibly assigning every input to a food or beverage category. The generated

Table 6: Step-wise time and space complexity of FRGP.

| Step | Time Complexity | Space Complexity |
|---|---|---|
| Exact semantic similarity / kNN graph construction | $O(n^2 d)$ | $O(nd + nK)$ |
| MST construction on the sparse kNN graph | $O(nK \log n)$ | $O(nK)$ |
| One-round probabilistic perturbation and component extraction | $O(n)$ | $O(n)$ |
| One-round spectral clustering on sampled subgraph $S^{(\ell)}$ | $O(tkm_\ell + s_\ell k^2 I)$ | $O(m_\ell + s_\ell k)$ |
| One-round Hungarian label alignment | $O(k^3)$ | $O(k^2)$ |
| Assignment aggregation over $L$ rounds | $O(Ln)$ | $O(nk)$ |
| Final entropy computation | $O(nk)$ | $O(nk)$ |

descriptions are subsequently encoded using the acge-text-embedding model to obtain the semantic embeddings used by FRGP.

**Ambiguity-scoring prompt.** For the LLM baseline, we query DeepSeek V3.2 using the following prompt, where "example dish name" is replaced with the input dish name: "We aim to identify ambiguous samples among food-delivery menu items. A sample is considered ambiguous if it exhibits characteristics of two or more categories. For example, 'hamburger + cola' belongs to both the fast-food and beverage categories, while 'tomato-and-egg rice set' combines the rice and stir-fried-dish categories; both are therefore ambiguous. First analyze the food category or categories of 'example dish name', and then assign it an integer ambiguity score from 1 to 10, where 10 indicates a clearly ambiguous sample with high confidence and 1 indicates a sample that stably belongs to a single well-defined category."

## E IMPACT STATEMENT

The primary application of FRGP is content organization on online platforms (e.g., food delivery, e-commerce), where it may benefit category management and user experience. We do not anticipate negative societal consequences directly arising from this method. Potential limitations include: (1) sensitivity to different embedding models: the quality of semantic embeddings depends on the underlying language model, which may inherit biases present in its training data; (2) the cold-start problem for newly added texts remains to be further investigated.

## F DERIVATION OF THE TIME AND SPACE COMPLEXITIES

Let $n$ denote the number of samples, $d$ the embedding dimension, $K$ the kNN parameter, $k$ the number of clusters, and $L$ the number of perturbation rounds. Let $s_\ell$ and $m_\ell$ denote the number of nodes and edges in the sampled subgraph at round $\ell$, and let $\bar{s}$ and $\bar{m}$ denote their averages over $L$ rounds. Let $t$ denote the eigensolver iteration count and $I$ the k-means iteration count.

In principle, the semantic similarity structure induced by pairwise sample relations is dense. Therefore, we choose dense full-graph spectral clustering as the baseline for complexity comparison. FRGP reduces this dense structure into a sparse computational pipeline by first constructing a $K$-nearest-neighbor graph and then extracting an MST. This sparsification is important both computationally and methodologically: it reduces memory usage and concentrates structural fragility near cluster boundaries.

Table 6 summarizes the time and space complexity of each step in FRGP.

A brief explanation of each term is given below. Under exact cosine kNN search, computing pairwise similarities costs $O(n^2 d)$. Since each node keeps only $K$ neighbors, the resulting sparse graph has $O(nK)$ edges, and constructing the MST on this sparse graph costs $O(nK \log n)$. At each perturbation round, cutting edges on the MST and extracting connected components costs $O(n)$, since the MST contains only $n - 1$ edges. Spectral clustering is then performed on the sampled subgraph. For round $\ell$, constructing the normalized Laplacian costs $O(m_\ell)$ and is absorbed into the sparse eigensolver cost $O(tkm_\ell)$. The subsequent k-means step is applied to $s_\ell$ embedded points in a $k$-dimensional spectral space and assigns them to $k$ centers; hence its cost is $O(s_\ell k^2 I)$. Therefore,

Table 7: Complexity comparison between FRGP and dense full-graph spectral clustering.

| Method | Time Complexity | Space Complexity |
|---|---|---|
| Dense full-graph spectral clustering | $O(n^2d + n^3)$ | $O(nd + n^2)$ |
| FRGP | $O\big(n^2d + nK\log n + L\left(n + tk\bar{m} + \bar{s}k^2I + k^3\right)\big)$ | $O(nd + nK + nk)$ |

the clustering cost at round $\ell$ is

$$O(tkm_\ell + s_\ell k^2 I).$$

After clustering, FRGP aligns the cluster labels to a reference partition using the Hungarian algorithm on a $k \times k$ cost matrix, which adds $O(k^3)$ per round. Therefore, the total cost of round $\ell$ is

$$O\big(n + tkm_\ell + s_\ell k^2 I + k^3\big).$$

Summing over $L$ rounds gives

$$O\big(L\left(n + tk\bar{m} + \bar{s}k^2 I + k^3\right)\big).$$

Finally, FRGP aggregates cluster assignments across the $L$ rounds and computes a normalized entropy score for each sample. With a streaming implementation that updates assignment counts online, this step costs $O(Ln)$ time, while the final entropy computation costs $O(nk)$ and is lower-order.

Combining the above steps, the overall time complexity of FRGP is

$$O\big(n^2d + nK\log n + L\left(n + tk\bar{m} + \bar{s}k^2 I + k^3\right)\big).$$

Combining the above steps, the working memory complexity is

$$O(nd + nK + nk).$$

Here, $O(nd)$ corresponds to storing embeddings, $O(nK)$ to the sparse kNN graph, with MST storage absorbed into the same order, and $O(nk)$ to the per-sample assignment counters. Since these statistics can be accumulated online, it is not necessary to store all $L$ clustering outputs explicitly.

We compare FRGP with a baseline that performs single-run spectral clustering directly on a dense full graph. Under a standard dense implementation, the baseline shares the same graph-construction cost $O(n^2d)$, then performs one full-graph spectral decomposition, which is typically upper bounded by $O(n^3)$. Its memory complexity is typically $O(nd+n^2)$, since it requires storing embeddings and a dense $n \times n$ similarity or Laplacian matrix.

Table 7 summarizes the comparison.

This comparison highlights two key points. First, **FRGP achieves lower space complexity.** FRGP does not operate on the original dense semantic graph throughout the whole pipeline; instead, it sparsifies the structure through kNN and MST. Therefore, when $K \ll n$ and $k \ll n$, its space complexity $O(nd + nK + nk)$ is substantially lower than the $O(nd + n^2)$ space complexity of dense full-graph spectral clustering. Second, **FRGP reduces the spectral-clustering cost.** Although FRGP repeats clustering over $L$ perturbed subgraphs, each round is performed on a sampled subgraph rather than the full graph. When the sampled subgraph size satisfies $\bar{s} \ll n$ and $k$ is moderate, the per-round spectral clustering cost is much smaller than a full-graph spectral decomposition. Thus, even with repeated perturbation rounds, FRGP can be computationally competitive or more efficient than dense full-graph spectral clustering in practice.

In practice, the FRGP method is first applied offline on historical data to identify ambiguous samples and generate stable labels or pseudo-labels, after which a lightweight parametric model can be trained for online deployment.