LLMs Meet GNNs for Authorship Attribution: Metric Geometry and Optimal-Transport Calibration

Anonymous authors
Paper under double-blind review

Abstract

Authorship attribution has long relied on function words, which capture syntactic style independently of topic and anonymize the text. Word adjacency networks (WANs) model each text as a Markov chain over function words, and distances between them define a syntactic geometry over a corpus. We exploit this geometry with graph neural networks (GNNs), which process large language model (LLM) embeddings of the anonymized texts as signals over a graph built from WAN distances. GNNs are stable and transferable across geometric graphs, but these guarantees hold only if the graph itself is well-behaved. We therefore design the graph so that its distances form a valid metric, remain stable under small edits, and keep distinct styles separated, and so that its edge weights make informative edges stand out. We characterize how candidate metrics respond to small perturbations, trace a token-level edit through every stage of the pipeline to prove stability and discriminability, and introduce an optional calibration step that reshapes the edge weights through a monotone, bi-Lipschitz optimal transport map. On Early Modern English plays, the calibrated GNN reaches 73.2% accuracy, outperforming an LLM classifier at 70.0% and an uncalibrated GNN at 70.5%, and calibration reduces the accuracy drop under WAN perturbations from 7.3% to 0.3%.

1 Introduction

Authorship attribution is the task of identifying the author of a text among a set of candidates. It underlies forensic linguistics, plagiarism detection, and the study of disputed works (Stamatatos, 2009; Juola, 2008). Since the study of the Federalist Papers by Mosteller & Wallace (1963), a central line of work has relied on function words, the articles, prepositions, pronouns, and conjunctions that carry grammatical structure rather than topical meaning. Authors use them unconsciously and consistently across topics, which makes them a stable stylistic signature (Koppel et al., 2009). Relying on function words alone also anonymizes the text, which matters when documents are private, sensitive, or partially redacted. Word adjacency networks (WANs) capture this signature by modeling each text as a Markov chain over function words, where transitions record how often one function word follows another (Segarra et al., 2015; 2016). Distances between WANs thus reflect how two texts are written rather than what they are about.

Large language models (LLMs) are the de facto standard for representing text, and their embeddings capture sequential patterns even in anonymized text (Staab et al., 2024; Varshney et al., 2020; Weggenmann et al., 2022; Carlini et al., 2021). However, they embed each text in isolation, so a classifier trained on these embeddings ignores the relationships across texts in a corpus. WAN distances provide exactly these relationships. They define a *corpus geometry* in which texts with similar syntactic habits lie close together. This suggests treating LLM embeddings as signals over a graph built from the syntactic geometry of WANs, so that author classification can exploit both.

Graph neural networks (GNNs) (Kipf & Welling, 2016; Veličković et al., 2017; Hamilton et al., 2017; Xu et al., 2018; Ruiz et al., 2021) are the natural tool for exploiting this geometry. They are permutation equivariant (Maron et al., 2018), stable to perturbations of the graph (Gama et al., 2020), and crucially, transferable across geometric graphs sampled from the same manifold (Ruiz et al., 2020). In our setting, these properties are essential, as WANs are estimated from finite,

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.

possibly noisy texts, and corpora grow as new texts are added. These guarantees, however, only hold if the graph itself is well-behaved. The distances across WANs thus must:

(O1) form a valid metric;
(O2) remain stable under small edits to the texts;
(O3) keep distinct styles separated;
(O4) yield edge weights distributed so that informative edges stand out.

Establishing objectives **(O1)**–**(O4)** is the focus of this paper.

We propose a three-stage pipeline, summarized in Figure 1, that meets these conditions by construction. In the first stage, we reduce each text to its function words, build its WAN, and embed the anonymized text with a pretrained LLM. In the second stage, we compute distances between the stationary distributions of the WANs, choosing a divergence that is a valid metric, and build a sparse k-nearest-neighbor graph whose edge weights are given by a bi-Lipschitz kernel of these distances. In the third stage, a GNN processes the LLM embeddings as signals over this graph to classify authorship. Between the second and third stages, an optional calibration step reshapes the edge weights. It learns a sparse reference graph from the LLM embeddings (Kalofolias, 2016) and maps the WAN edge weights onto the reference weight distribution through a one-dimensional optimal transport (OT) map, which preserves the rank order of the edges. Our contributions are as follows.

(C1) We characterize how candidate metrics respond to small perturbations, showing they partition into three sensitivity families (Fisher information, ℓ_1 norm, or sub-linear), guiding divergence selection for **(O1)** and **(O2)**.
(C2) We trace a token-level edit through every stage of the pipeline and prove stability and discriminability at each stage, meeting goals **(O2)** and **(O3)**.
(C3) We show that OT calibration addresses **(O4)** while preserving **(O1)**–**(O3)**, as the transport map is monotone and bi-Lipschitz on its active support.
(C4) On 773 text segments from Early Modern English plays, the calibrated GNN reaches 73.2% accuracy, outperforming both an LLM classifier at 70.0% and the uncalibrated GNN at 70.5%. The gain over the LLM classifier shows that the syntactic geometry of WANs adds information beyond the embeddings alone, and calibration further reduces the accuracy drop under WAN perturbations from 7.3% to 0.3%.

Section 2 describes the pipeline, Sections 3–6 develop its guarantees stage by stage, and Section 7 reports numerical results.

2 Method Overview

Authorship attribution relies on stylistic signatures that persist across topics. Syntactic habits are one such signature, as authors repeat them unconsciously regardless of what they write about (Mosteller & Wallace, 1963; Stamatatos, 2009; Segarra et al., 2015). Relying on syntax alone also anonymizes the text, which matters when documents are private, sensitive, or partially redacted. We capture syntax through function words, the closed class of words such as articles, prepositions, pronouns, conjunctions, and auxiliary verbs that carry grammatical structure rather than topical meaning. We reduce each text to its function words and replace every other word type with a placeholder, s.a., the word entity type or PoS tag. We then process these anonymized texts through a modular three-stage design, summarized in Figure 1 and detailed below.

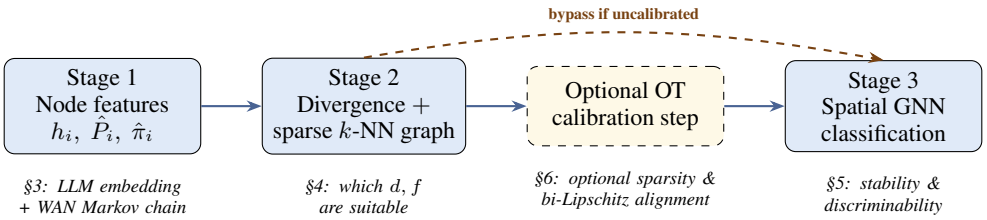

Figure 1: The LLM-GNN pipeline with optional OT calibration and corresponding theory sections.

Stage 1: Node Feature and WAN Extraction. Let $X = \{x_1, \ldots, x_N\}$ be a corpus of N anonymized texts. Each text x_i is mapped to (a) an (unsupervised) LLM embedding $h_i \in \mathbb{R}^d$ from a pretrained language model, and (b) a so-called WAN. Originally defined by Segarra et al. (2015), WANs are graph representations of writing style for authorship attribution. They describe how an

author combines *function words* (e.g., conjunctions, prepositions, and determiners), i.e., words that primarily express grammatical relations. The premise is that function words carry little topical content, so their co-occurrence patterns reflect an author's style more than the subject. Formally, let $\mathcal{F} = \{f_1, \ldots, f_N\}$ be a fixed, ordered set of N function words, let $\alpha \in (0, 1)$ be a discount factor, and let $D \geq 1$ be an integer window length. A text x is split into H sentences delimited by stopper symbols (period, question mark, exclamation mark, and semicolon), and $w_{h,e}$ denotes the word at position e of sentence h, which has L_h words. The WAN of x is the weighted directed graph $\mathcal{G}_x = (\mathcal{F}, \mathcal{E}_x, \boldsymbol{W}_x)$ with edge set $\mathcal{E}_x = \{(f_i, f_j) : [\boldsymbol{W}_x]_{ij} > 0\}$ and weighted adjacency matrix $\boldsymbol{W}_x \in \mathbb{R}_{\geq 0}^{N \times N}$ given by

$$[\boldsymbol{W}_x]_{ij} = \sum_{h=1}^{H} \sum_{\substack{1 \leq e < e' \leq L_h \\ e' - e \leq D}} \alpha^{e'-e-1} \, \mathbf{1}\{w_{h,e} = f_i, \; w_{h,e'} = f_j\}. \tag{1}$$

In words, each occurrence of f_j located $d \leq D$ positions after an occurrence of f_i in the same sentence adds α^{d-1} to the weight of the edge (f_i, f_j).

To make WANs of texts with different lengths comparable, each row of $\boldsymbol{W}_x$ is normalized by the out-strength $s_{x,i} = [\boldsymbol{W}_x \mathbf{1}]_i$, which gives the row-stochastic matrix $\boldsymbol{P}_x$ with entries

$$[\boldsymbol{P}_x]_{ij} = [\boldsymbol{W}_x]_{ij} / s_{t,i} \;\text{ if } s_{t,i} > 0, \qquad [\boldsymbol{P}_x]_{ij} = 1/N \;\text{ if } s_{x,i} = 0. \tag{2}$$

The matrix $\boldsymbol{P}_x$ is the transition matrix of a Markov chain on $\mathcal{F}$, assumed ergodic (Segarra et al., 2015), which holds for sufficiently long texts. Ergodicity guarantees a unique stationary distribution $\boldsymbol{\pi}_x$ with $\boldsymbol{P}_x^k \to \mathbf{1}\boldsymbol{\pi}_x^\top$ as $k \to \infty$.

Stage 2: Metric Construction and Sparsification. Pairwise divergences $d(\hat{\pi}_i, \hat{\pi}_j)$ between stationary distributions are computed for all pairs of texts to build a sparse k-NN graph $G = (\mathcal{V}, \mathcal{E})$. This leaves us with a weight distribution μ that consists of a Dirac mass at 0 and an active component μ_{active}. A monotone kernel κ maps distances to raw edge weights, $c_{ij} = \kappa(d(\hat{\pi}_i, \hat{\pi}_j))$. Two design questions arise at this stage, both of which we analyze mathematically in Section 4: which divergence d is best suited to computing WAN distances (Sections 4.1-4.2), and what regularity must κ satisfy to avoid amplifying noise or collapsing distinct texts (Section 4.3).

Stage 3: Spatial Graph Convolution. A GNN performs message passing over $(G, \{h_1, \ldots, h_N\})$ to produce per-text authorship logits. Since Stage 2 turns the space of texts into a metric space, we can view the texts as samples from a manifold embedded in it and G as a geometric graph built from these samples. Extensive prior work on geometric deep learning (Bronstein et al., 2017; 2021) has shown that GNNs on geometric graphs such as G converge to manifold neural networks as N grows. This makes them transferable across graphs sampled from the same manifold and yields generalization bounds that depend on the manifold rather than on the particular sample (Wang et al., 2024a;b; 2025; Netto et al., 2025a;b). In our case, we expect this to mean that a GNN trained on one set of texts remains reliable when new texts are added or when it is applied to a different sample from the same authors. Section 5 further derives a discriminability guarantee for this architecture when the activation has a strictly positive derivative, such as LeakyReLU (Theorem 5.2). Empirically, both LLM+WAN variants outperform the single-modality baselines—LLM+MLP and GCN (Table 1)—confirming that the graph-based fusion of semantic and syntactic representations captures complementary stylistic information.

Topological Calibration via Optimal Transport (Optional). While the Stage 2 graph can be fed directly to the GNN, fixing a metric gives no control over the distribution of the resulting edge weights. If the weights concentrate sharply around their mean, the graph behaves as a nearly complete graph with uniform weights. If the weights spread out as in a uniform distribution, informative and uninformative edges become hard to separate. Since the most informative edges are those with the largest weights, a distribution that separates high-weight edges from the bulk is preferable, as it also allows sparsifying the graph for computational efficiency. To reshape the weight distribution accordingly, we introduce an optional calibration step.

More specifically, we calibrate the raw graph's edge weights against a reference distribution derived from the LLM embeddings $H = [h_1, \ldots, h_N]^\top$. We learn a reference graph from these embeddings using the formulation of Kalofolias (2016), which yields a sparse, connected adjacency matrix W^* whose properties are characterized Corollary 6.1. The empirical distribution of the entries of W^* serves as the reference distribution ν, a mixture of a Dirac mass at 0 and an active component ν_{active}.

Mapping the empirical CDF F_μ of the active WAN weights μ_{active} onto the CDF F_ν of ν_{active} via the monotone transport map $f = F_\nu^{-1} \circ F_\mu$ yields calibrated weights $w_{ij} = f(c_{ij})$, preserving the exact WAN-derived rank ordering of edges Theorem 6.2.

Robustness Requirements. Together, these stages map each anonymized text to a node in a metric graph, optionally calibrate its edge weights, and classify authorship through message passing over LLM embeddings. For this pipeline to be reliable, however, each stage must behave well under the noise inherent in finite texts, e.g., from limited sample length, or from transcription and editorial errors (Belinkov & Bisk, 2017; Pruthi et al., 2019; Buys & Blunsom, 2017).

WANs are empirical estimates computed from finite texts, so small changes to a text, such as light edits or transcription errors, perturb them. Each stage of the pipeline transforms these estimates, and any stage that amplifies small perturbations can turn them into spurious graph structure. Stability is also a central reason to use GNNs in the first place. Beyond permutation equivariance, GNNs are provably stable to graph perturbations and transferable across graphs (Ruiz et al., 2021), but these guarantee only help if the graph itself is estimated stably from the texts. Stability alone is not enough, since a map that sends every text to the same output is trivially stable. We therefore also require discriminability, so that texts with different WANs remain separated. Formally, stability requires Lipschitz continuity (bounding output deviation), while discriminability requires a lower bi-Lipschitz bound (preserving separation). Sections 3–5 establish both properties stage by stage, tracing a token-level edit through every stage up to the GNN output.

3 Stage 1: Stationary Distributions

We start by examining how word-level edits to a text perturb its WAN and, in turn, its stationary distribution, and whether texts by different authors remain separated at this stage.

Let $P, \tilde{P}$ be the transition matrices of the unperturbed and perturbed WAN over vocabulary V, with stationary distributions $\pi, \tilde{\pi}$ ($\pi^\top P = \pi^\top$; $\tilde{\pi}^\top \tilde{P} = \tilde{\pi}^\top$), and let $A = I - P$. Since $A\mathbf{1} = 0$, A is singular with $\pi^\top A = 0$; let $A^\sharp$ be its group inverse, satisfying $AA^\sharp = A^\sharp A = I - \mathbf{1}\pi^\top$ (Meyer, 1975). Theorem 3.1 establishes a stability result on the perturbation of the stationary distribution under local vocabulary edits:

Theorem 3.1 (Perturbation of the Stationary Distribution)**.** $\|\tilde{\pi} - \pi\|_1 \leq \|\tilde{P} - P\|_\infty \|A^\sharp\|_1$. Because a text edit alters at most Δ raw transition counts across C distinct vocabulary rows in a document of length L, the perturbation is bounded as $\|\tilde{P} - P\|_\infty \leq C\Delta/L$, which implies that $\|\tilde{\pi} - \pi\|_1 \leq (C\Delta/L)\|A^\sharp\|_1$.

Proof. See Appendix B.5. □

Theorem 3.1 decouples the sensitivity of π to token edits into a local edit magnitude $C\Delta/L$ and a global structural factor $\|A^\sharp\|_1$, where $A^\sharp = (I - P)^\sharp$ is the group inverse of the discrete Markov Laplacian. Topologically, $\|A^\sharp\|_1$ tracks mixing time: fast-mixing chains with large spectral gaps $\gamma = 1 - \lambda_2(P)$ dissipate perturbations exponentially, while slow-mixing or bottlenecked topologies amplify local $O(L^{-1})$ edits into systemic shifts in π.

To establish the corresponding discriminability result, we first formalize the notion of an *effective syntax shift* (i.e., the condition under which two authorial representations are structurally distinct).

Definition 3.1 (Effective syntax shift)**.** Authors A, B are *effectively distinct* if $\pi_A^\top P_A \neq \pi_A^\top P_B$ or $\pi_B^\top P_A \neq \pi_B^\top P_B$.

The syntactic shift $\delta_{A\to B} = \|\pi_A^\top (P_A - P_B)\|_1$ measures the expected discrepancy when Author B's transition operator P_B acts on Author A's equilibrium distribution π_A. This shift is inherently asymmetric ($\delta_{A\to B} \neq \delta_{B\to A}$), capturing how one author's phrase construction triggers syntactic anomalies against another's stylistic habits, providing a word-usage-based anchor for authorial separability. Theorem 3.2 establishes that stationary divergence cannot vanish without the underlying syntactic shift also collapsing.

Theorem 3.2 (Discriminability of Stationary Distributions)**.** For any two authors A and B, $\|\pi_A - \pi_B\|_1 \geq \frac{1}{2}\max\{\delta_{A\to B}, \delta_{B\to A}\}$.

Proof. See Appendix B.6. □

By lower-bounding the ℓ_1 distance between stationary distributions by half the effective syntax shift, the theorem ensures that structural discrepancies in word-pairing habits induce measurable separation in equilibrium profiles. Distinct authorial styles thus remain provably distinguishable, preventing disparate transition dynamics from collapsing into indistinguishable signatures.

4 Stage 2: The Metric Graph

We next examine how perturbations of the stationary distributions propagate to the pairwise distances and edge weights of the metric graph.

4.1 Sensitivity and Asymptotic Hierarchy of Divergences

Let $\Delta^{n-1} = \{\pi \in \mathbb{R}^n : \pi_i \geq 0,\ \sum_i \pi_i = 1\}$ denote the probability simplex; all distributions are assumed strictly positive where logarithms appear. We consider five standard candidate divergences that can be used as metrics on Δ^{n-1}: total variation distance TV, Hellinger distance d_H, Jensen–Shannon divergence JS, Fisher–Rao distance FR, and k-Wasserstein distance W_k ($k \geq 1$), whose ground metric is the normalized shortest-path distance between function words (Appendix D.1.5). Formal definitions are provided in Appendix B.1.

To guide divergence selection, we characterize how each metric responds to a small perturbation of a stationary distribution—the event triggered by a text edit at Section 3.

Theorem 4.1 (Asymptotic Hierarchy Under Local Perturbation)**.** Let $p \in \operatorname{int}(\Delta^{n-1})$ and $q = p + \varepsilon v$ with $\sum_i v_i = 0$, $\|v\|_\infty \leq C$, $q \in \Delta^{n-1}$. Let $\mathcal{I}_F = \sum_i v_i^2/p_i$ (the Fisher information along v). Then, as $\varepsilon \to 0$,

$$d_{\mathrm{TV}}(p,q) = \tfrac{\varepsilon}{2}\|v\|_1, \qquad d_H(p,q) = \tfrac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F} + O(\varepsilon^2),$$
$$d_{\mathrm{JS}}(p,q) = \tfrac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F} + O(\varepsilon^2), \qquad d_{\mathrm{FR}}(p,q) = \varepsilon\sqrt{\mathcal{I}_F} + O(\varepsilon^2),$$

and, on a finite discrete space, $W_k(p,q) = \Theta\big(\varepsilon^{1/k}\|v\|_1^{1/k}\big)$ for every $k \geq 1$.

Consequently these quantities decouple into three families:

1. **linear w.r.t. Fisher information**, $d_H \sim d_{\mathrm{JS}} \sim \frac{1}{2\sqrt{2}} d_{\mathrm{FR}} \sim \frac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F}$;
2. **linear w.r.t. perturbations**, $2d_{\mathrm{TV}} \sim W_1 \sim \varepsilon\|v\|_1$;
3. **sub-linear Wasserstein**, $W_k \sim \varepsilon^{1/k}\|v\|_1^{1/k}$, $k \geq 1$.

Proof. See Appendix B.1. □

Theorem 4.1 shows that the candidate metrics respond to a small perturbation εv in three distinct ways. The distances d_H, d_{JS}, and d_{FR} grow with $\sqrt{\mathcal{I}_F} = \sqrt{\sum_i v_i^2/p_i}$, which weights each change by $1/p_i$. A shift in a rare transition therefore counts more than the same shift in a common one. This makes these metrics sensitive to rare stylistic habits, but rare transitions are also the most poorly estimated from finite texts. The distances d_{TV} and W_1 instead grow with $\|v\|_1$ and do not depend on p, so every unit of moved mass counts the same regardless of how frequent the transition is. Finally, W_k with $k > 1$ grows as $\varepsilon^{1/k}$, which is much larger than ε for small ε. Small perturbations thus produce comparatively large distances, which makes W_k sensitive to subtle stylistic shifts but also to estimation noise. Section D.2 empirically validates these predictions: the measured scaling exponents match theory, and the Fisher-information-based metrics are approximately 2.6 times more sensitive to rare-word perturbations than d_{TV} or W_1 (Table 2).

4.2 Stability of Pairwise Distances

Suppose the stationary distributions are perturbed such that $\tilde{\pi}_i = \pi_i + \delta_i$ and $\tilde{\pi}_j = \pi_j + \delta_j$, where $\mathbf{1}^\top \delta_i = 0$ and $\mathbf{1}^\top \delta_j = 0$. For any metric d on Δ^{V-1}, the reverse triangle inequality gives

$$|d(\tilde{\pi}_i, \tilde{\pi}_j) - d(\pi_i, \pi_j)| \leq d(\pi_i, \tilde{\pi}_i) + d(\pi_j, \tilde{\pi}_j), \tag{3}$$

the proof of which is provided in Appendix B.7. Here, the deviation in distance between two perturbed stationary distributions is upper-bounded by the sum of their individual perturbation distance,

guaranteeing metric stability under noises. For any metric introduced in Section 4.1, we refer to Theorem 4.1 for the derivation of the individual perturbation distance $d(\pi_i, \tilde{\pi}_i)$. *We skip the discriminability result here since $\pi_i, \pi_j, \tilde{\pi}_i, \tilde{\pi}_j$ can be distinct and need not force $d(\tilde{\pi}_i, \tilde{\pi}_j)$ and $d(\pi_i, \pi_j)$ to be different.*

4.3 Kernel Regularity

To ensure stability and discriminability in this layer, we require the kernel κ mapping distances to initial edge weights to be **bi-Lipschitz**. Specifically, there exist constants $c, C > 0$ such that

$$c|d_1 - d_2| \leq |\kappa(d_1) - \kappa(d_2)| \leq C|d_1 - d_2|$$

for all $d_1, d_2 \in [d_{\min}, d_{\max}]$, where $d_{\max} < \infty$ bounds the chosen metric on the simplex and $d_{\min} \geq 0$ is the smallest distance in the graph. Since every metric in Section 4.1 is bounded, the regularized inverse kernel $\kappa(d) = (d + \alpha)^{-1}$ satisfies this condition with $d_{\min} = 0$. The Gaussian kernel $\kappa(d) = \exp(-d^2/2\sigma^2)$ satisfies it only for $d_{\min} > 0$, since its derivative vanishes at zero.

Here, the kernel κ simultaneously delivers **noise insulation** through its upper bound L_κ and an **anti-collapse guarantee** through its lower bound c. The upper Lipschitz bound $C = \sup_d |\kappa'(d)|$ acts as a geometric dampener, ensuring that estimation noise or micro-scale perturbations in stationary distributions δ_i cannot be artificially amplified into massive, destabilizing jumps in graph adjacency weights prior to GNN message passing. Conversely, the lower bound $c = \inf_d |\kappa'(d)| > 0$ enforces strict injectivity across the kernel's domain, preventing representation collapse by guaranteeing that distinct topological distances $d_1 \neq d_2$ map to strictly separable edge weights without flattening or compressing the underlying graph geometry.

5 Stage 3: GNN Outputs

Finally, we examine how perturbations of the graph and of the LLM embeddings propagate to the GNN outputs, and whether distinct inputs remain separated after message passing.

Let $Z = \sigma(WH\Theta)$ be the hidden representation matrix with input weight operator W (either the Stage 2 adjacency, or its calibrated version, see Section 6), embedding matrix H, model parameters Θ, and γ-Lipschitz activation σ (e.g. LeakyReLU with $\gamma = 1$). Let $\tilde{Z} = \sigma(\tilde{W}H\Theta)$ be the representation under the perturbed $\tilde{W}$. Under the weight perturbations $\Delta_W = \tilde{W} - W$ (and similarly the signal perturbations $\Delta_H = \tilde{H} - H$), we obtain the following stability result on the GNN output:

Theorem 5.1 (Stability on GNN Outputs)**.** $\|\tilde{Z} - Z\|_F \leq \gamma\|\Theta\|_2\|H\|_2\|\Delta_W\|_F$ and, symmetrically, $\|\tilde{Z} - Z\|_F \leq \gamma\|W\|_2\|\Theta\|_2\|\Delta_H\|_F$.

Proof. See Appendix B.8. □

Since Sections 4 and 6 bound individual edge perturbations $|\Delta W_{ij}|$, the Frobenius norm follows by summing over edges: $\|\Delta W\|_F \leq \sqrt{|E|}\max_{ij}|\Delta W_{ij}|$. Then the stability bound in Theorem 5.1 shows that the output error $\|\tilde{Z} - Z\|_F$ scales linearly with perturbations in either factor ($\|\Delta_W\|_F$ or $\|\Delta_H\|_F$), governed by the activation slope γ and spectral norms of the unperturbed operators. The structural symmetry between the two bounds confirms that feature-space and weight-space noise propagate identically. Consequently, constraining $\|\Theta\|_2$—for instance, via spectral normalization—guarantees robustness against structural perturbations without restricting model capacity.

To guarantee that structural perturbations produce distinct output embeddings, Theorem 5.2 leverages entrywise monotonicity in non-linear activations to lower-bound output divergence as a function of perturbation magnitude.

Theorem 5.2 (Discriminability on GNN Outputs)**.** If $\sigma' \geq \gamma_{\min} > 0$ on the relevant domain (true of LeakyReLU globally, and of ELU/GELU on bounded feature ranges, but false of ReLU, sigmoid, and tanh outside a compact window), then $\|\tilde{Z} - Z\|_F \geq \gamma_{\min}\sigma_{\min}(H)\sigma_{\min}(\Theta)\|\Delta_W\|_F$ and $\|\tilde{Z} - Z\|_F \geq \gamma_{\min}\sigma_{\min}(W)\sigma_{\min}(\Theta)\|\Delta_H\|_F$.

Proof. See Appendix B.9. □

Theorem 5.2 guarantees against feature collapse: output distance scales with the minimal singular values $\sigma_{\min}(H)$ and $\sigma_{\min}(\Theta)$. The presence of $\gamma_{\min}$ highlights the role of activation selection—

strictly monotonic activations like LeakyReLU preserve discriminability ($\gamma_{\min} > 0$), whereas ReLU or saturated activations permit perturbations to fall within dead zones. Maintaining full-rank features and parameters is thus necessary to prevent rank collapse, justifying spectral regularization to ensure distinct topological signals remain resolvable.

6 CALIBRATION

When the calibration step is used, it sits between the kernel of Section 4.3 and the GNN of Section 5. In the following, we show that the OT map f is bi-Lipschitz with constants m and M. Therefore, it only rescales the stability and discriminability bounds of Stage 2 by these constants, and the guarantees of Section 5 continue to hold for the calibrated graph.

6.1 A SPARSE REFERENCE GRAPH

Let $X \in \mathbb{R}^{N\times d}$ collect the LLM embeddings and $Z_{ij} = \|x_i - x_j\|^2$ the pairwise squared distances. To learn a geometrically regularized reference graph from the LLM embeddings, we employ Kalofolias (2016)'s formulation under a *Dirichlet energy signal smoothness assumption*[1]:

$$\min_{W\in\mathcal{W}} \sum_{i,j} W_{ij}Z_{ij} - \alpha \sum_i \log\Big(\sum_j W_{ij}\Big) + \frac{\beta}{2}\|W\|_F^2, \tag{4}$$
$$\text{where } \mathcal{W} = \{W = W^\top \geq 0,\ \operatorname{diag}(W) = 0\};\ \alpha, \beta > 0.$$

The structural properties of the resulting adjacency matrix W^* are directly governed by its first-order stationarity. Theorem 6.1 establishes the element-wise KKT optimality equation that underpins all downstream topological characteristics of this learned reference graph.

Theorem 6.1 (Karush-Kuhn-Tucker (KKT) Optimality)**.** Let $d_i = \sum_{k\neq i} W_{ik}$ denote the degree of node i. Then any minimizer W^* satisfies, for $i < j$ and multiplier $\lambda_{ij} \geq 0$ with $\lambda_{ij}W_{ij} = 0$,

$$2Z_{ij} - \alpha\Big(\tfrac{1}{d_i} + \tfrac{1}{d_j}\Big) + 2\beta W_{ij} - \lambda_{ij} = 0.$$

Proof. See Appendix B.2. □

For the non-negativity inequality constraints $W^*_{ij} \geq 0$, the standard KKT complementary slackness condition mandates that $\lambda_{ij}W^*_{ij} = 0$ and $\lambda_{ij} \geq 0$, which splits the distribution of W^*_{ij} into two distinct geometric phases:

- **The Dirac mass at exact zero**: $W^*_{ij} = 0, \lambda_{ij} \geq 0 \Leftrightarrow Z_{ij} \geq \frac{\alpha}{2}\left(d_i^{-1} + d_j^{-1}\right)$. When $Z_{ij} > \frac{\alpha}{2}(d_i^{-1} + d_j^{-1})$, W_{ij} collapses to 0. The parameter α thus functions as a *structural thresholding driver*, determining the proportion of active edges ($W_{ij} > 0$) relative to zero-weight entries.
- **The active component**: $\lambda_{ij} = 0 \Leftrightarrow W^*_{ij} = \frac{\alpha}{2\beta}\left(d_i^{-1} + d_j^{-1}\right) - \frac{1}{\beta}Z_{ij} > 0$. Here, increasing α *shifts mass away from* 0 while increasing β *compresses the active range*.

Another property worth noting is that Kalofolias (2016) guarantees **the learned graph to be fully connected without isolated nodes**. Trivially from the objective function equation 4, the logarithmic degree barrier $-\alpha\sum_i \log d_i$ approaches $+\infty$ as any node degree $d_i \to 0^+$, ensuring that no minimizer W^* can send a node's degree strictly to zero. This connectivity guarantee ensures that every text remains reachable during message passing, preventing pathological cases where disconnected nodes receive no neighborhood signal and default to purely local (LLM) classification.

Corollary 6.1 (Sparsity, Tail Shape, and Connectivity)**.** *(i)* If $Z_{ij} \geq \frac{\alpha}{2}(d_i^{-1}+d_j^{-1})$ then $W^*_{ij} = 0$ exactly. *(ii)* Otherwise $W^*_{ij} = \frac{\alpha}{2\beta}(d_i^{-1} + d_j^{-1}) - \frac{1}{\beta}Z_{ij} \in \big(0,\ \frac{\alpha}{2\beta}(d_i^{-1} + d_j^{-1})\big]$. *(iii)* No node is isolated ($d_i > 0$ at any minimizer).

We summarize the properties of the learned graph in Corollary 6.1. The following proposition characterizes when the active component in (ii) is right-skewed.

[1]Though our LLM embeddings do not necessarily satisfy this assumption, smoothness can be enforced via regularization approaches such as weight decay (Belkin et al., 2006; Zhou & Schölkopf, 2005; Smola & Kondor, 2003), which are commonly employed in the training of deep learning models.

Proposition 6.1 (Right-skewness of Active Weights)**.** Assume approximately homogeneous degrees $d_i \approx \bar{d}$ and that the density p_Z of Z_{ij} is nondecreasing near zero. Then the density p_W of active weights $W^*_{ij} > 0$ is nonincreasing, hence right-skewed.

Proof. See Appendix B.10. □

For LLM embeddings with intrinsic dimension D, the density of squared pairwise distances near zero scales as $p_Z(z) \propto z^{D/2-1}$, which is nondecreasing whenever $D \geq 2$—a condition easily satisfied by modern sentence embeddings. By Proposition 6.1, this yields a right-skewed distribution of active weights. This encodes precisely the prior that most text pairs carry no stylistic signal while a sparse subset exhibits meaningful similarity—the foundational assumption behind thresholding-based graph sparsification. A uniform target distribution would offer no natural cutoff; a tight symmetric target would be unstable, with small threshold changes shifting substantial mass. The right-skewed shape near zero with a thin active tail yields stable, interpretable sparsity, following the same logic as the disparity filter for network backbones (Serrano et al., 2009) and the graphical lasso's Laplace prior (Friedman et al., 2008).

6.2 A Monotone Bi-Lipschitz Transport Map

Let p_μ and p_ν denote the PDFs of the active raw WAN-kernel weights μ_{active} and of ν_{active}, respectively, and Let F_μ and F_ν denote the corresponding CDFs. The transport map is $f(c) = F_\nu^{-1}(F_\mu(c))$, whose key properties are provided in Theorem 6.2.

Theorem 6.2 (Structural Properties of f)**.** Assume p_μ, p_ν are strictly positive and continuous. Then f is (a) monotone (rank-preserving), $c_1 \leq c_2 \Rightarrow f(c_1) \leq f(c_2)$; (b) piecewise differentiable with $f'(c) = p_\mu(c)/p_\nu(f(c))$; and (c) bi-Lipschitz on any compact active sub-domain $\mathcal{C}$, $m|c_1 - c_2| \leq |f(c_1) - f(c_2)| \leq M|c_1 - c_2|$ for $0 < m \leq M < \infty$.

Proof. See Appendix B.4. □

Theorem 6.2 establishes f as a regularized homeomorphic embedding underpinning the pipeline's end-to-end guarantees. The target ν_{active}—an active component on $(0, W_{\max}]$ with $W_{\max} = \max_{i,j} \frac{\alpha}{2\beta}(d_i^{-1} + d_j^{-1})$—applies a nonlinear elasticity that compresses low-similarity connections near zero and expands high-similarity pairs into a compact tail. Regardless of how noisy the empirical distribution $F_\mu(c)$ remains, active weights are mapped into bounded support, ensuring the downstream GNN operates on a graph that is Lipschitz-continuous with respect to perturbations, informative with respect to authorial style, and stable during training.

7 Experiments

Experimental Setup. We evaluate LLM-GNN on 154 plays by six Early Modern English playwrights (Segarra et al., 2015), segmented into approximately 5,000-word chunks to yield 773 graph nodes (majority-class baseline: 0.308). Each chunk is content-masked by replacing content words with POS or entity-type tags while retaining function words; the same masked text is used for both WAN construction and LLM encoding (`all-MiniLM-L6-v2`). Unless otherwise stated, the k-NN graph ($k = 3$) uses Hellinger distances with edge weights $\kappa(d) = (1+d)^{-1}$; for OT calibration, the reference graph is learned via the Kalofolias objective with $\alpha = \beta = 1$. We report mean $\pm$ standard deviation over 20 random splits for authorship attribution and 10 random splits for auxiliary tasks, with paired t-tests for statistical comparisons. Full details are in Appendix D.

Table 1: Test accuracy (mean $\pm$ standard deviation) on authorship attribution and auxiliary tasks.

Method	Authorship	Location	Time Period
GCN-only	0.614 ± 0.032	0.794 ± 0.064	0.716 ± 0.031
LLM-MLP	0.700 ± 0.047	0.868 ± 0.044	0.898 ± 0.030
LLM-GCN	0.705 ± 0.028	$\mathbf{0.926 \pm 0.083}$	0.872 ± 0.032
LLM-GCN-OT	$\mathbf{0.732 \pm 0.043}$	0.901 ± 0.052	$\mathbf{0.915 \pm 0.028}$

Main Authorship Classification Results. We compare a GCN-only baseline, LLM-MLP, and LLM-GCN with raw or OT-calibrated WAN weights (Table 1). The OT-calibrated model achieves 0.732 mean accuracy, outperforming both LLM-MLP (0.700, $p = 0.0006$) and raw-weight LLM-GCN (0.705, $p = 0.014$). The gain from OT calibration arises from reshaping edge weight distributions: raw weights may concentrate around their mean, causing over-smoothing, whereas calibration

introduces a sparse informative tail (Section 6) that amplifies contrast between stylistically similar and dissimilar pairs. The gap between GCN-only (0.614) and LLM-based methods confirms that semantic features are essential, while improvement over LLM-MLP demonstrates that WAN-derived syntactic geometry provides complementary information.

Location and Time Period Classification. We further evaluate on two auxiliary tasks: author location (American vs. English; 345 nodes) and time period (Early Modern vs. Victorian vs. Modern; 608 nodes). The best-performing model on both tasks combines LLM and WAN information, though the optimal weighting differs: raw-weight LLM-GCN excels on location (0.926 vs. 0.868 for LLM-MLP), while OT calibration yields the best time-period accuracy (0.915 vs. 0.898 for LLM-MLP and 0.872 for raw-weight). This difference likely reflects signal granularity: geographic distinctions manifest in coarse syntactic patterns well-captured by raw WAN distances, whereas temporal distinctions involve subtler stylistic shifts that benefit from calibration's enhanced edge contrast.

Robustness to WAN Perturbations. We test the stability guarantees of Section 5 by perturbing each WAN transition matrix as $\tilde{P} = (1-\varepsilon)\hat{P} + \varepsilon R$, where R is a random row-stochastic matrix, increasing ε from 0 to 0.5 while keeping LLM features fixed (see Appendix D.1.8 for further details on experimental setup). As shown in Figure 2, accuracy degrades gradually with increasing noise, consistent with the theoretical stability bounds. OT calibration substantially reduces this sensitivity: the accuracy drop from $\varepsilon = 0$ to 0.5 decreases from 0.073 (raw) to 0.003 (calibrated), averaged across six divergences. This robustness is consistent with the bi-Lipschitz property of the calibration map (Section 6), which preserves the Stage 3 stability guarantees while reshaping the weight distribution.

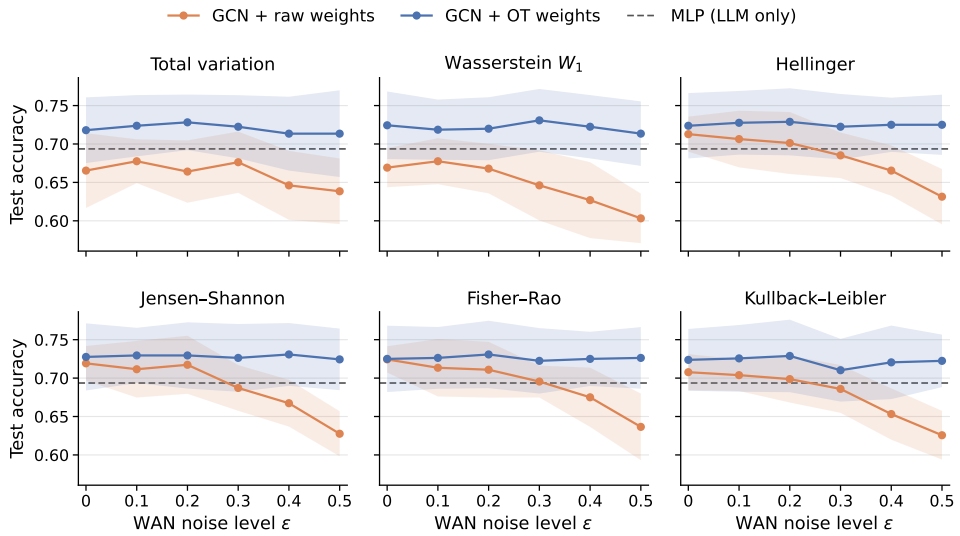

Figure 2: Test accuracy under increasing WAN perturbation for raw and OT-calibrated edge weights across six divergences. Shaded regions: ± 1 standard deviation over 10 seeds; dashed line: MLP baseline.

8 Conclusion

We presented a framework for authorship attribution that fuses LLM embeddings with the syntactic geometry of word adjacency networks through graph neural networks. Our theoretical contributions establish stability and discriminability guarantees at each pipeline stage, characterize how candidate divergences respond to perturbations, and show that OT calibration reshapes edge weights via a bi-Lipschitz map that preserves these guarantees while amplifying informative edges.

Empirically, the calibrated GNN achieves 73.2% accuracy on Early Modern English plays versus 70.0% (LLM-only) and 70.5% (uncalibrated), confirming that WAN geometry provides complementary information. OT calibration reduces accuracy drop under perturbations from 7.3% to 0.3%, demonstrating robustness for short or noisy texts. Future work includes learning task-specific vocabularies, scaling via approximate nearest-neighbor methods, and extending to other stylistic tasks.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for grammar and spelling checks. We have not used generative AI tools for writing core technical content, generating proofs, or producing experimental results, and tasks such as data collection and human subject research are not applicable to this work. Additionally, we used generative AI tools for improving clarity and conciseness of exposition. We have reviewed all AI-assisted work; all suggested edits were manually verified by the authors for correctness and consistency with our intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Reproducibility Statement All theoretical results include complete proofs in Appendix **??**, with assumptions clearly stated in the main text. Experimental details including dataset preprocessing, WAN construction, graph building, model architectures, and training hyperparameters are provided in Appendix D. The primary authorship corpus is from Segarra et al. (2015), and auxiliary corpora are obtained from Project Gutenberg (`https://www.gutenberg.org`). Anonymous source code implementing the full pipeline is available at `https://anonymous.4open.science/r/LLM-fixed-GNN-trained-Authorship-Attribution-ECF8`.

REFERENCES

Nicholas Andrews and Marcus Bishop. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1684–1695, 2019.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7 (11), 2006.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL `https://arxiv.org/abs/2104.13478`.

Jan Buys and Phil Blunsom. Robust incremental neural semantic graph parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1215–1226, 2017.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Patrick Juola. Authorship attribution for electronic documents. In *IFIP International Conference on Digital Forensics*, pp. 119–130. Springer, 2006.

Patrick Juola. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3): 233–334, 2008.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial intelligence and statistics*, pp. 920–929. PMLR, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Carl D Meyer, Jr. The role of the group generalized inverse in the theory of finite markov chains. *Siam Review*, 17(3):443–464, 1975.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

Caio F. Deberaldini Netto, Zhiyang Wang, and Luana Ruiz. Improved image classification with manifold neural networks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025a. doi: 10.1109/ICASSP49660.2025.10889415.

Caio F. Deberaldini Netto, Zhiyang Wang, and Luana Ruiz, 2025b. URL https://arxiv.org/abs/2506.12197.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

G Peyré and M Cuturi. Computational optimal transport. found trends mach learn 11 (5–6): 355–607, 2019.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5582–5591, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Rafael A Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 913–919, 2021.

Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33:1702–1712, 2020.

Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. Graph neural networks: Architectures, stability, and transferability. *Proceedings of the IEEE*, 109(5):660–682, 2021.

Santiago Segarra, Mark Eisen, and Alejandro Ribeiro. Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, 63(20):5464–5478, 2015.

Santiago Segarra, Mark Eisen, Gabriel Egan, and Alejandro Ribeiro. Attributing the authorship of the henry vi plays by word adjacency. *Shakespeare Quarterly*, 67(2):232–256, 2016.

M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16): 6483–6488, 2009.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning theory and kernel machines: 16th annual conference on learning theory and 7th kernel workshop, COLT/kernel 2003, Washington, DC, USA, august 24-27, 2003. Proceedings*, pp. 144–158. Springer, 2003.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations*, volume 2024, pp. 33832–33878, 2024.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

Lav R Varshney, Nitish Shirish Keskar, and Richard Socher. Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)*, pp. 1–5. IEEE, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Z. Wang, L. Ruiz, and A. Ribeiro. Geometric graph filters and neural networks: Limit properties and discriminability trade-offs. 2024a.

Zhiyang Wang, Juan Cervino, and Alejandro Ribeiro. Generalization of Geometric Graph Neural Networks. *Asilomar Conference on Signals, Systems, and Computers*, 2024b.

Zhiyang Wang, Juan Cervino, and Alejandro Ribeiro. A manifold perspective on the statistical generalization of graph neural networks, 2025. URL https://arxiv.org/abs/2406.05225.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, pp. 721–731, 2022.

Anna Wegmann and Dong Nguyen. Does it capture stel? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7109–7130, 2021.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pp. 361–368. Springer, 2005.

A Appendix

A Related Work

Authorship Attribution and Stylometry. Quantitative authorship attribution originated with Mosteller & Wallace (1963)'s study of the *Federalist Papers*, establishing that unconsciously used function words—rather than topic-heavy content words—provide a stable authorial signature. This principle, surveyed by Juola (2006) and Koppel et al. (2009), guides the field: features must capture *how* rather than *what* is written to ensure cross-genre generalization and avoid topic bias. Subsequent work cataloged vector-based features (n-grams, POS patterns) paired with Bayesian or SVM classifiers (Stamatatos, 2009). Shifted from feature vectors to network topology, Segarra et al. (2015) introduced Word Adjacency Networks (WANs), modeling texts as function-word Markov chains and comparing authors via spectral graph distances. Our Layer-1 stability analysis (Section **??**) formalizes and proves the foundational, previously unproven intuition that WANs are robust to incidental lexical variation.

Neural and LLM-Based Approaches to Style. NLP document representations shifted from static word vectors (Mikolov et al., 2013; Pennington et al., 2014) to dense neural embeddings like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), USE (Cer et al., 2018), and Sentence-BERT (Reimers & Gurevych, 2019), both designed primarily for semantic similarity rather than style. Consequently, work has focused on topic-invariant author representations (Andrews & Bishop, 2019), contrastive fine-tuning for universal authorship embeddings (Rivera-Soto et al., 2021), and content-controlled evaluation frameworks (Wegmann & Nguyen, 2021; Wegmann et al., 2022) to untangle style from topic. However, existing approaches fine-tune or probe embeddings as black boxes without offering a geometric account of their interaction with structural representations like WANs.

Combining off-the-shelf embeddings with WANs introduces an unstudied commensurability problem between semantic manifolds and syntactic network geometries, motivating our theoretical analysis in Sections 4.1-6.2.

Graph Signal Processing and Learning Graphs from Data. Inferring graph topology from data is central to graph signal processing (Shuman et al., 2013; Ortega et al., 2018), which leverages signal smoothness to constrain admissible topologies. Dong et al. (2016) and Kalofolias (2016) formulate graph learning as convex optimization, using Dirichlet energy smoothness alongside log-barrier degree penalties and Frobenius regularizers to enforce connectivity and bound the weight magnitude. We adopt the Kalofolias formulation for topological calibration because its KKT conditions (Theorem 6.1) yield explicit, monotone hyperparameter control over sparsity (Corollary 6.1), providing a principled reference distribution to calibrate an independently constructed graph, a application unexplored in prior work.

OT for Distribution Alignment and Calibration. OT provides a principled framework for comparing and transforming probability distributions (Villani et al., 2009), popularized in machine learning via entropic regularization (Cuturi, 2013; Peyré & Cuturi, 2019; Kolouri et al., 2017) and domain adaptation (Courty et al., 2016). In the context of multi-modal fusion, OT offers robust properties against metric mismatch and supports stable geometry preservation even when supports differ or overlap imperfectly (Gabriel & Marco, 2019). Building on these foundations, we leverage monotone quantile matching (Villani et al., 2009) to align two edge-weight distributions from different modalities—semantic embeddings and syntactic WAN statistics—over the same text set. This adopts the degree-regularized topology of the embedding graph while preserving WAN stylistic rankings.

B Proofs

B.1 Proof of Theorem 4.1

We begin by recalling the formal definitions of seven standard divergences used to measure distributional discrepancies.

Definition B.1 (Divergence Metrics on Stationary Distributions)**.** Let $p, q \in \Delta^{n-1}$ be probability distributions over n states, and let $\mathrm{BC}(p,q) = \sum_i \sqrt{p_i q_i}$ denote the Bhattacharyya coefficient. We define the seven candidate WAN divergences as follows:

Kullback–Leibler (KL) Divergence:

$$D_{\mathrm{KL}}(p\|q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right).$$

Cross-Entropy:

$$H(p\|q) = -\sum_{i=1}^{n} p_i \log q_i = H(p) + D_{\mathrm{KL}}(p\|q),$$

where $H(p) = -\sum_{i=1}^{n} p_i \log p_i$.

Total Variation (TV) Distance:

$$\mathrm{TV}(p,q) = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i| = \max_{A\subseteq[n]} |p(A) - q(A)|.$$

Hellinger Distance:

$$d_H(p,q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2 = \sqrt{1 - \mathrm{BC}(p,q)}.$$

Jensen–Shannon Divergence (JSD) & Metric (JS)**:**

$$\mathrm{JS}(p,q) = \sqrt{\mathrm{JSD}(p,q)},$$
$$\text{where } \mathrm{JSD}(p,q) = \frac{1}{2}D_{\mathrm{KL}}(p\|m) + \frac{1}{2}D_{\mathrm{KL}}(q\|m),\ m = \frac{1}{2}(p+q).$$

Fisher–Rao Distance:

$$\mathrm{FR}(p,q) = 2\arccos\left(\sum_{i=1}^{n}\sqrt{p_i q_i}\right) = 2\arccos\big(\mathrm{BC}(p,q)\big).$$

k-Wasserstein Distance ($k \geq 1$):

$$W_k(p,q) = \left(\inf_{\gamma\in\Pi(p,q)}\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^k\, d\gamma(x,y)\right)^{1/k},$$

where $\Pi(p,q)$ is the set of valid couplings on metric space $(\mathcal{X}, d)$.

We now prove each asymptotic expansion in turn.

Kullback–Leibler Divergence and Cross-Entropy. Since $p \in \mathrm{int}(\Delta^{n-1})$, we can safely Taylor-expand $\log(p_i + \varepsilon v_i)$ around p_i for sufficiently small ε:

$$\log(p_i + \varepsilon v_i) = \log p_i + \varepsilon\frac{v_i}{p_i} - \frac{\varepsilon^2}{2}\frac{v_i^2}{p_i^2} + O(\varepsilon^3)$$

Substituting this into Definition 1.1 yields:

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= \sum_{i=1}^{n} p_i \log\frac{p_i}{p_i+\varepsilon v_i} = \sum_{i=1}^{n} p_i\left(\log p_i - \log(p_i+\varepsilon v_i)\right) \\ &= \sum_{i=1}^{n} p_i\left(-\varepsilon\frac{v_i}{p_i} + \frac{\varepsilon^2}{2}\frac{v_i^2}{p_i^2} + O(\varepsilon^3)\right) \\ &= -\varepsilon\underbrace{\sum_{i=1}^{n} v_i}_{0} + \frac{\varepsilon^2}{2}\sum_{i=1}^{n}\frac{v_i^2}{p_i} + O(\varepsilon^3) = \frac{\varepsilon^2}{2}\mathcal{I}_F + O(\varepsilon^3). \end{aligned}$$

Total Variation Distance. Direct evaluation from Definition 1.3 yields the exact linear relation without an error residue:

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2}\sum_{i=1}^{n}|p_i - (p_i+\varepsilon v_i)| = \frac{\varepsilon}{2}\sum_{i=1}^{n}|v_i| = \frac{\varepsilon}{2}\|v\|_1.$$

Hellinger Distance. Expanding the Bhattacharyya coefficient to trace the exact error propagation:

$$\mathrm{BC}(p,q) = \sum_{i=1}^{n} p_i\sqrt{1+\varepsilon\frac{v_i}{p_i}} = \sum_{i=1}^{n} p_i\left(1 + \frac{\varepsilon v_i}{2p_i} - \frac{\varepsilon^2 v_i^2}{8p_i^2} + O(\varepsilon^3)\right)$$

Applying $\sum_{i=1}^{n} v_i = 0$ simplifies this to

$$\mathrm{BC}(p,q) = 1 - \frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3). \tag{5}$$

From Definition 1.4,

$$d_{\mathrm{H}}(p,q)^2 = 1 - \mathrm{BC}(p,q) = \frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3).$$

Taking the principal square root mapping with the standard Taylor identity $\sqrt{1+O(\varepsilon)} = 1 + O(\varepsilon)$ yields:

$$d_{\mathrm{H}}(p,q) = \sqrt{\frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3)} = \frac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F}\sqrt{1+O(\varepsilon)} = \frac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F} + O(\varepsilon^2).$$

Jensen–Shannon Distance. By Definition 1.5, the squared Jensen–Shannon distance is given by the divergence $\mathrm{JSD}(p,q) = \frac{1}{2}D_{\mathrm{KL}}(p \parallel m) + \frac{1}{2}D_{\mathrm{KL}}(q \parallel m)$, where the midpoint distribution is $m = \frac{p+q}{2} = p + \frac{\varepsilon}{2}v$.

We evaluate each Kullback–Leibler divergence term individually. First, for $D_{\mathrm{KL}}(p \parallel m)$, we Taylor-expand $\log(m_i) = \log(p_i + \frac{\varepsilon}{2}v_i)$ around p_i:

$$\log\left(p_i + \frac{\varepsilon}{2}v_i\right) = \log p_i + \frac{\varepsilon}{2}\frac{v_i}{p_i} - \frac{\varepsilon^2}{8}\frac{v_i^2}{p_i^2} + O(\varepsilon^3)$$

Substituting this expansion into the definition of $D_{\mathrm{KL}}(p \parallel m)$ yields:

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel m) &= \sum_{i=1}^{n} p_i \left(\log p_i - \log m_i\right) = \sum_{i=1}^{n} p_i \left(-\frac{\varepsilon}{2}\frac{v_i}{p_i} + \frac{\varepsilon^2}{8}\frac{v_i^2}{p_i^2} + O(\varepsilon^3)\right) \\ &= -\frac{\varepsilon}{2}\underbrace{\sum_{i=1}^{n} v_i}_{0} + \frac{\varepsilon^2}{8}\sum_{i=1}^{n}\frac{v_i^2}{p_i} + O(\varepsilon^3) = \frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3). \end{aligned}$$

Second, for $D_{\mathrm{KL}}(q \parallel m)$, we must expand both the base measure $q = p + \varepsilon v$ and the mixture $m = p + \frac{\varepsilon}{2}v$ simultaneously:

$$D_{\mathrm{KL}}(q \parallel m) = \sum_{i=1}^{n}(p_i + \varepsilon v_i)\log\left(\frac{p_i + \varepsilon v_i}{p_i + \frac{\varepsilon}{2}v_i}\right) = \sum_{i=1}^{n}(p_i + \varepsilon v_i)\log\left(1 + \frac{\frac{\varepsilon}{2}v_i}{p_i + \frac{\varepsilon}{2}v_i}\right)$$

Using the standard logarithmic expansion $\log(1+x) = x - \frac{x^2}{2} + O(x^3)$ evaluated at $x = \frac{\frac{\varepsilon}{2}v_i}{p_i + \frac{\varepsilon}{2}v_i}$ and the standard geometric series expansion $(1+u)^{-1} = 1 - u + u^2 - O(u^3)$, we obtain:

$$x = \frac{\varepsilon v_i}{2p_i}\left(1 + \frac{\varepsilon v_i}{2p_i}\right)^{-1} = \frac{\varepsilon v_i}{2p_i} - \frac{\varepsilon^2 v_i^2}{4p_i^2} + O(\varepsilon^3) \quad \text{and} \quad \frac{x^2}{2} = \frac{\varepsilon^2 v_i^2}{8p_i^2} + O(\varepsilon^3).$$

Subtracting these yields the expansion of the interior log term:

$$\log\left(1 + \frac{\frac{\varepsilon}{2}v_i}{p_i + \frac{\varepsilon}{2}v_i}\right) = \frac{\varepsilon v_i}{2p_i} - \frac{3\varepsilon^2 v_i^2}{8p_i^2} + O(\varepsilon^3)$$

Distributing across the leading factor $(p_i + \varepsilon v_i)$ inside the summation:

$$\begin{aligned} D_{\mathrm{KL}}(q \parallel m) &= \sum_{i=1}^{n}(p_i + \varepsilon v_i)\left(\frac{\varepsilon v_i}{2p_i} - \frac{3\varepsilon^2 v_i^2}{8p_i^2} + O(\varepsilon^3)\right) \\ &= \sum_{i=1}^{n}\left(\frac{\varepsilon v_i}{2} - \frac{3\varepsilon^2 v_i^2}{8p_i} + \frac{\varepsilon^2 v_i^2}{2p_i} + O(\varepsilon^3)\right) \\ &= \frac{\varepsilon}{2}\underbrace{\sum_{i=1}^{n} v_i}_{0} + \frac{\varepsilon^2}{8}\sum_{i=1}^{n}\frac{v_i^2}{p_i} + O(\varepsilon^3) = \frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3). \end{aligned}$$

Averaging the two identically scaled symmetric expansions reveals the total divergence:

$$\mathrm{JSD}(p,q) = \frac{1}{2}\left(\frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3)\right) + \frac{1}{2}\left(\frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3)\right) = \frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3).$$

To compute the final metric distance $d_{\text{JS}}(p,q) = \sqrt{\text{JSD}(p,q)}$, we extract the dominant quadratic parameter out from under the radical and use the binomial identity $\sqrt{1+O(\varepsilon)} = 1 + O(\varepsilon)$:

$$d_{\text{JS}}(p,q) = \sqrt{\frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3)} = \frac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F}\sqrt{1+O(\varepsilon)} = \frac{\varepsilon}{2\sqrt{2}}\sqrt{\mathcal{I}_F} + O(\varepsilon^2).$$

Fisher–Rao Distance. From Definition 1.6, $d_{\text{FR}}(p,q) = 2\arccos(\text{BC}(p,q))$. Using equation 5 and the expansion $\arccos(1-x) = \sqrt{2x} + O(x^{3/2})$ with the standard Taylor identity $\sqrt{1+O(\varepsilon)} = 1 + O(\varepsilon)$:

$$\begin{aligned} d_{\text{FR}}(p,q) &= 2\sqrt{2\left(\frac{\varepsilon^2}{8}\mathcal{I}_F + O(\varepsilon^3)\right)} + O(\varepsilon^{3/2}) \\ &= 2\left(\frac{\varepsilon}{2}\sqrt{\mathcal{I}_F}\right)\sqrt{1+O(\varepsilon)} + O(\varepsilon^{3/2}) \\ &= \varepsilon\sqrt{\mathcal{I}_F} + O(\varepsilon^2). \end{aligned}$$

Wasserstein Distance. Let $\mathcal{X} = \{1,\dots,n\}$ be a finite metric space equipped with a metric d. Let $\Pi(p,q)$ denote the set of all valid coupling matrices (joint distributions) on $\mathcal{X}\times\mathcal{X}$ with marginal distributions p and q. The k-th Wasserstein distance is defined as:

$$W_k(p,q) = \left(\inf_{\gamma'\in\Pi(p,q)} \sum_{i,j=1}^{n} d(i,j)^k \gamma'_{ij}\right)^{1/k} = \left(\inf_{\gamma'\in\Pi(p,q)} \sum_{i\neq j} d(i,j)^k \gamma'_{ij}\right)^{1/k} \tag{6}$$

where the second equality follows from the identity $d(i,i) = 0$ for all $i \in \mathcal{X}$.

Let $D = \max_{i,j} d(i,j) < \infty$ and $c = \min_{i\neq j} d(i,j) > 0$. For any valid coupling $\gamma' \in \Pi(p,q)$, bounding the ground cost coordinates yields $c^k \le d(i,j)^k \le D^k$ for all $i \neq j$. Monotonicity of the summation implies:

$$c^k \sum_{i\neq j}\gamma'_{ij} \le \sum_{i\neq j} d(i,j)^k \gamma'_{ij} \le D^k \sum_{i\neq j}\gamma'_{ij} \tag{7}$$

To bound the off-diagonal mass $\sum_{i\neq j}\gamma'_{ij}$, we leverage the properties of Total Variation distance. For any $\gamma' \in \Pi(p,q)$ and any index i, we observe that $\gamma'_{ii} \le \sum_{j=1}^n \gamma'_{ij} = p_i$ and $\gamma'_{ii} \le \sum_{j=1}^n \gamma'_{ji} = q_i$, which implies $\gamma'_{ii} \le \min(p_i, q_i)$. Consequently, the total mass preserved on the diagonal for *any* feasible coupling is globally upper-bounded:

$$\sum_{i=1}^{n}\gamma'_{ii} \le \sum_{i=1}^{n}\min(p_i,q_i)$$

It follows that the off-diagonal mass for any arbitrary coupling $\gamma' \in \Pi(p,q)$ satisfies:

$$\sum_{i\neq j}\gamma'_{ij} = 1 - \sum_{i=1}^{n}\gamma'_{ii} \ge 1 - \sum_{i=1}^{n}\min(p_i,q_i)$$

Using the algebraic identity $\min(p_i,q_i) = \frac{1}{2}(p_i+q_i-|p_i-q_i|)$ and the fact that $\sum_i p_i = \sum_i q_i = 1$, we evaluate this off-diagonal mass explicitly:

$$\begin{aligned} 1 - \sum_{i=1}^{n}\min(p_i,q_i) &= 1 - \sum_{i=1}^{n}\frac{p_i+q_i-|p_i-q_i|}{2} \\ &= 1 - \left(\frac{1}{2}\sum_{i=1}^{n}p_i + \frac{1}{2}\sum_{i=1}^{n}q_i - \frac{1}{2}\sum_{i=1}^{n}|p_i-q_i|\right) \\ &= 1 - (1 - d_{\text{TV}}(p,q)) = d_{\text{TV}}(p,q). \end{aligned}$$

Applying the left inequality of equation 7 to the definition equation 6:

$$W_k(p,q) \geq \left(\inf_{\gamma' \in \Pi(p,q)} c^k \sum_{i \neq j} \gamma'_{ij} \right)^{1/k} \geq \left(c^k \cdot d_{\text{TV}}(p,q) \right)^{1/k} = c \cdot d_{\text{TV}}(p,q)^{1/k}$$

Next, we construct a specific candidate coupling $\gamma \in \Pi(p,q)$ to attain the upper bound. By the definition of the maximal coupling, we set the diagonal entries to $\gamma_{ii} = \min(p_i, q_i)$. The remaining residual mass can be distributed across the off-diagonal entries via the product of the marginal residuals: $\gamma_{ij} = \frac{(p_i - \min(p_i,q_i))(q_j - \min(p_j,q_j))}{d_{\text{TV}}(p,q)}$ for $i \neq j$ (if $d_{\text{TV}}(p,q) = 0$, then $p = q$ and $\gamma_{ij} = 0$ trivially). By construction, this specific coupling satisfies $\sum_{i \neq j} \gamma_{ij} = 1 - \sum_{i=1}^{n} \min(p_i, q_i)$. Furthermore in Appendix B.3, we prove that

$$\sum_{i \neq j} \gamma_{ij} = d_{\text{TV}}(p,q). \tag{8}$$

Since $W_k(p,q)$ minimizes over the feasible set, evaluating the cost at our specific constructed coupling γ yields an upper bound:

$$W_k(p,q) \leq \left(\sum_{i \neq j} d(i,j)^k \gamma_{ij} \right)^{1/k} \leq \left(D^k \sum_{i \neq j} \gamma_{ij} \right)^{1/k} = D \cdot d_{\text{TV}}(p,q)^{1/k}$$

Substituting the structural perturbation identity $d_{\text{TV}}(p,q) = \frac{\varepsilon}{2}\|v\|_1$ into these inequalities yields the final rigorous result:

$$c \left(\frac{\varepsilon}{2} \|v\|_1 \right)^{1/k} \leq W_k(p,q) \leq D \left(\frac{\varepsilon}{2} \|v\|_1 \right)^{1/k}$$

B.2 Proof of Theorem 6.1

By symmetrizing the objective over the upper triangle, as $W_{ij} = W_{ji} \geq 0$ by the symmetry of the adjacency matrix, we rewrite the objective function as the Lagrangian

$$\mathcal{L}(W, \Lambda) = 2 \sum_{i<j} W_{ij} Z_{ij} - \alpha \sum_{i=1}^{N} \log \left(\sum_{k \neq i} W_{ik} \right) + \beta \sum_{i<j} W_{ij}^2 - \sum_{i<j} \lambda_{ij} W_{ij}$$

Taking the partial derivative of the Lagrangian $\mathcal{L}$ with respect to a specific independent edge variable W_{ij} (where $i < j$) and setting it to zero yields:

$$\frac{\partial \mathcal{L}}{\partial W_{ij}} = 2Z_{ij} - \frac{\alpha}{\sum_{k \neq i} W_{ik}} - \frac{\alpha}{\sum_{k \neq j} W_{jk}} + 2\beta W_{ij} - \lambda_{ij} = 0$$

Let $d_i = \sum_{k \neq i} W_{ik}$ represent the degree of node i. The first-order optimality condition simplifies directly to:

$$2Z_{ij} - \alpha \left(\frac{1}{d_i} + \frac{1}{d_j} \right) + 2\beta W_{ij} - \lambda_{ij} = 0$$

B.3 Proof of Equation 8

Let $u_i = p_i - \min(p_i, q_i)$ be the residual mass of p at index i, and let $v_j = q_j - \min(p_j, q_j)$ be the residual mass of q at index j.

Step 1: Summing the Residuals Separately

Before evaluating the joint sum, we establish the individual sums of u_i and v_j over all coordinates. Using the identity $1 - \sum_{i=1}^{n} \min(p_i, q_i) = d_{\text{TV}}(p,q)$ and the total mass conservation property

($\sum_{i=1}^{n} p_i = 1$), the total sum of u_i is evaluated as:

$$\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} (p_i - \min(p_i, q_i)) = \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} \min(p_i, q_i)$$
$$= 1 - \sum_{i=1}^{n} \min(p_i, q_i) = d_{\text{TV}}(p, q)$$

By symmetric logic, noting that $\sum_{j=1}^{n} q_j = 1$, summing the v_j residuals yields an identical total:

$$\sum_{j=1}^{n} v_j = \sum_{j=1}^{n} (q_j - \min(p_j, q_j)) = \sum_{j=1}^{n} q_j - \sum_{j=1}^{n} \min(p_j, q_j)$$
$$= 1 - \sum_{j=1}^{n} \min(p_j, q_j) = d_{\text{TV}}(p, q)$$

Thus, the sum of each individual residual vector is exactly $d_{\text{TV}}(p, q)$.

STEP 2: EVALUATING THE DOUBLE SUMMATION

We now sum the off-diagonal coupling formula over all pairs $i \neq j$.

Note on indices: If $i = j$, then either $u_i = 0$ or $v_i = 0$ because $\min(p_i, q_i)$ must equal either p_i or q_i, which nullifies one of the terms. Consequently, the product $u_i v_i = 0$ for all diagonal elements. This means summing over $i \neq j$ is algebraically equivalent to summing over all $i, j \in \{1, \ldots, n\}$.

Plugging the residual definitions into the joint sum yields:

$$\sum_{i \neq j} \gamma_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{u_i \cdot v_j}{d_{\text{TV}}(p, q)} = \frac{1}{d_{\text{TV}}(p, q)} \sum_{i=1}^{n} \sum_{j=1}^{n} (u_i \cdot v_j)$$
$$= \frac{1}{d_{\text{TV}}(p, q)} \sum_{i=1}^{n} \left(u_i \cdot \sum_{j=1}^{n} v_j \right) = \frac{1}{d_{\text{TV}}(p, q)} \sum_{i=1}^{n} (u_i \cdot d_{\text{TV}}(p, q))$$
$$= \sum_{i=1}^{n} u_i = d_{\text{TV}}(p, q).$$

B.4 PROOF OF THEOREM 6.2

(a) is a standard property of the monotone (Monge) transport map between one-dimensional distributions.

(b) Since p_μ and p_ν are strictly positive and continuous, the CDFs F_μ and F_ν are continuously differentiable. Thus, by the Inverse Function Theorem, $f(c)$ is continuously differentiable on the active domain.

Since $f(c) = F_\nu^{-1}(F_\mu(c))$, $F_\nu(f(c)) = F_\mu(c)$. Hence,

$$\frac{d}{dc} \left[F_\nu(f(c)) \right] = \frac{d}{dc} \left[F_\mu(c) \right] = p_\mu(c).$$

Note that

$$\frac{d(\text{Outer})}{d(f(c))} = p_\nu(f(c)) \quad \text{and} \quad \frac{d(\text{Inner})}{dc} = \frac{df(c)}{dc}$$

Using the chain rule, we multiply the derivatives together

$$\frac{d}{dc} \left[F_\nu(f(c)) \right] = p_\nu(f(c)) \cdot \frac{df(c)}{dc},$$

and so

$$\frac{df(c)}{dc} = \frac{p_\mu(c)}{p_\nu(f(c))}$$

(c) Since p_μ and p_ν are continuous and strictly positive on their respective compact domains, they achieve strictly positive lower bounds and finite upper bounds by the Extreme Value Theorem. Thus, there exist constants $0 < c_{\min} \le c_{\max} < \infty$ such that for all raw weights $c \in \mathcal{C}$:

$$c_{\min} \le p_\mu(c) \le c_{\max} \quad \text{and} \quad c_{\min} \le p_\nu(f(c)) \le c_{\max}.$$

Substituting these bounds into the derivative expression yields:

$$0 < m = \frac{c_{\min}}{c_{\max}} \le f'(c) \le \frac{c_{\max}}{c_{\min}} = M < \infty.$$

For any two weights $c_1, c_2 \in \mathcal{C}$ with $c_1 \neq c_2$, since f is continuously differentiable on the active domain, we can apply the Mean Value Theorem. There exists some intermediate point ξ between c_1 and c_2 such that:

$$f(c_1) - f(c_2) = f'(\xi)(c_1 - c_2).$$

Taking the absolute value on both sides and applying our derivative bounds $m \le f'(\xi) \le M$:

$$m|c_1 - c_2| \le |f(c_1) - f(c_2)| \le M|c_1 - c_2|.$$

B.5 Proof of Theorem 3.1

Using $\tilde{\pi}^\top \mathbf{1} = 1$ and $AA^\sharp = I - \mathbf{1}\pi^\top$, we obtain the identity

$$\tilde{\pi}^\top - \pi^\top = \tilde{\pi}^\top I - (\tilde{\pi}^\top \mathbf{1})\pi^\top = \tilde{\pi}^\top AA^\sharp = (\tilde{\pi}^\top - \tilde{\pi}^\top P)A^\sharp = \tilde{\pi}^\top(\tilde{P} - P)A^\sharp.$$

Taking ℓ_1 norms and the compatibility of induced matrix norms,

$$\begin{aligned}\|\tilde{\pi} - \pi\|_1 &= \|\tilde{\pi}^\top(\tilde{P} - P)A^\sharp\|_1 \le \|\tilde{\pi}\|_1 \|\tilde{P} - P\|_\infty \|A^\sharp\|_1 \\ &= \|\tilde{P} - P\|_\infty \|A^\sharp\|_1 \le (C\Delta/L)\|A^\sharp\|_1.\end{aligned}$$

B.6 Proof of Theorem 3.2

Using $\pi_B^\top A_B = \mathbf{0}^\top$,

$$\pi_A^\top(P_A - P_B) = \pi_A^\top - \pi_A^\top P_B = \pi_A^\top A_B = (\pi_A^\top - \pi_B^\top)A_B.$$

Applying $\|x^\top M\|_1 \le \|x^\top\|_1 \|M\|_\infty$ to the right-hand side produces an upper bound based on the distribution gap:

$$\|\pi_A^\top(P_A - P_B)\|_1 = \|(\pi_A^\top - \pi_B^\top)A_B\|_1 \le \|\pi_A - \pi_B\|_1 \|A_B\|_\infty.$$

Because P_B is a row-stochastic matrix,

$$\|A_B\|_\infty = \|I - P_B\|_\infty \le \|I\|_\infty + \|P_B\|_\infty = 2.$$

Hence, $\|\pi_A^\top(P_A - P_B)\|_1 \le 2\|\pi_A - \pi_B\|_1$, which implies $\|\pi_A - \pi_B\|_1 \ge \frac{1}{2}\|\pi_A^\top(P_A - P_B)\|_1 = \frac{1}{2}\delta_{A\to B}$. Combining this with $\|\pi_A - \pi_B\|_1 \ge \frac{1}{2}\delta_{B\to A}$ obtained via an identical argument yields $z\|\pi_A - \pi_B\|_1 \ge \frac{1}{2}\max\{\delta_{A\to B}, \delta_{B\to A}\}$.

B.7 Proof of Equation 3

Step 1: Establishing the Upper Bound

We expand the perturbed distance $d(\tilde{\pi}_i, \tilde{\pi}_j)$ by introducing a sequential detour through the unperturbed states π_i and π_j. Applying the triangle inequality twice yields:

$$d(\tilde{\pi}_i, \tilde{\pi}_j) \le d(\tilde{\pi}_i, \pi_i) + d(\pi_i, \pi_j) + d(\pi_j, \tilde{\pi}_j)$$

Invoking metric symmetry to substitute $d(\tilde{\pi}_i, \pi_i) = d(\pi_i, \tilde{\pi}_i)$, and subtracting $d(\pi_i, \pi_j)$ from both sides of the inequality, we isolate the direct distance variation:

$$d(\tilde{\pi}_i, \tilde{\pi}_j) - d(\pi_i, \pi_j) \le d(\pi_i, \tilde{\pi}_i) + d(\pi_j, \tilde{\pi}_j) \tag{9}$$

STEP 2: ESTABLISHING THE LOWER BOUND

Symmetrically, we evaluate the unperturbed distance $d(\pi_i, \pi_j)$ by routing its path through the perturbed states $\tilde{\pi}_i$ and $\tilde{\pi}_j$:

$$d(\pi_i, \pi_j) \leq d(\pi_i, \tilde{\pi}_i) + d(\tilde{\pi}_i, \tilde{\pi}_j) + d(\tilde{\pi}_j, \pi_j)$$

Utilizing metric symmetry ($d(\tilde{\pi}_j, \pi_j) = d(\pi_j, \tilde{\pi}_j)$) and rearranging terms to isolate the same distance variation yields:

$$d(\pi_i, \pi_j) - d(\tilde{\pi}_i, \tilde{\pi}_j) \leq d(\pi_i, \tilde{\pi}_i) + d(\pi_j, \tilde{\pi}_j)$$

Multiplying the entire inequality by -1 reverses the inequality direction, establishing our formal lower bound:

$$d(\tilde{\pi}_i, \tilde{\pi}_j) - d(\pi_i, \pi_j) \geq -\left(d(\pi_i, \tilde{\pi}_i) + d(\pi_j, \tilde{\pi}_j)\right) \tag{10}$$

STEP 3: SYNTHESIS VIA ABSOLUTE VALUE DEFINITION

Combining the upper constraint from equation 9 and the lower constraint from equation 10, we obtain

$$|d(\tilde{\pi}_i, \tilde{\pi}_j) - d(\pi_i, \pi_j)| \leq d(\pi_i, \tilde{\pi}_i) + d(\pi_j, \tilde{\pi}_j)$$

B.8 PROOF OF THEOREM 5.1

We obtain

$$\|\tilde{Z} - Z\|_F = \|\sigma(\tilde{W}H\Theta) - \sigma(WH\Theta)\|_F \leq \gamma\|\Delta_W H\Theta\|_F \leq \gamma\|\Delta_W\|_F\|H\|_2\|\Theta\|_2$$

by sub-multiplicativity, and symmetrically for Δ_H.

B.9 PROOF OF THEOREM 5.1

Using $\|AB\|_F \geq \sigma_{\min}(A)\|B\|_F$ in place of sub-multiplicativity in the proof above, we have

$$\begin{aligned}\|\tilde{Z} - Z\|_F = \|\sigma(\tilde{W}H\Theta) - \sigma(WH\Theta)\|_F &\geq \gamma_{\min}\|\Delta_W H\Theta\|_F \\ &\geq \gamma_{\min}\sigma_{\min}(H)\sigma_{\min}(\Theta)\|\Delta_W\|_F,\end{aligned}$$

and symmetrically for Δ_H.

B.10 PROOF OF PROPOSITION 6.1

Under homogeneous degrees, $W^*_{ij} = c - \frac{1}{\beta}Z_{ij}$ for a constant c. By change of variables, $p_W(w) = \beta \cdot p_Z(\beta(c - w))$. Since this is a decreasing transformation and p_Z is nondecreasing near zero, p_W is nonincreasing on its support.

C LOCALIZED BI-LIPSCHITZ BOUNDS FOR OT CALIBRATION UNDER RIGHT-SKEWED TARGETS

The global bi-Lipschitz bounds established in Theorem 6.2(c) guarantee that the OT calibration map f neither collapses distinct weights (via the lower bound $m = \inf_c f'(c) > 0$) nor amplifies perturbations unboundedly (via the upper bound $M = \sup_c f'(c) < \infty$). However, due to the right-skewness of ν_{active}, these global bounds are necessarily loose: M must accommodate the sparse right tail where target density $p_\nu(f(c))$ is small, while m is constrained by the dense region near zero where $p_\nu(f(c))$ is large. This appendix derives localized, regime-specific bounds that provide tighter Lipschitz constants by restricting analysis to density-homogeneous subdomains.

First, we define localized, dynamic bounds constrained to the weight capacity $\mathcal{C}$:

$$\begin{aligned}m(\mathcal{C}) &= \inf_{c \geq \mathcal{C}} f'(c) = \inf_{c \geq \mathcal{C}} p_\mu(c)/p_\nu(f(c)); \\ M(\mathcal{C}) &= \sup_{c \leq \mathcal{C}} f'(c) = \sup_{c \leq \mathcal{C}} p_\mu(c)/p_\nu(f(c)).\end{aligned}$$

For a larger cutoff threshold $\mathcal{C}$, values $c \geq \mathcal{C}$ are pushed deeper into the sparse right tail to preserve rank order. This yields a small target density $p_\nu(f(c))$ and a correspondingly large derivative $f'(c)$. Consequently, the localized lower Lipschitz bound $m(\mathcal{C})$ tightens (increases), albeit at the cost of excluding the regime $c < \mathcal{C}$ where noise dominates. Conversely, a smaller threshold $\mathcal{C}$ captures $c \leq \mathcal{C}$ within the dense probability mass near zero, where $p_\nu(f(c))$ is large and $f'(c)$ is small. This decreases the localized upper Lipschitz bound $M(\mathcal{C})$, while giving up coverage for $c > \mathcal{C}$ where the data is sparse.

By establishing these localized bounds $m(\mathcal{C})$ and $M(\mathcal{C})$, we replace the unnecessarily loose global Lipschitz constant M with a regime-aware framework that directly accounts for target skewness. Practically, this refinement allows us to decoupled gradient stability analysis across different density regimes: the upper bound $M(\mathcal{C})$ provides a much tighter Lipschitz constant for the dense region near zero—where the bulk of the edge weights reside—thereby preventing conservative over-smoothing in message passing without being inflated by tail behavior. Conversely, the lower bound $m(\mathcal{C})$ isolates the sparse right tail ($c \geq \mathcal{C}$), guaranteeing a strictly positive derivative floor that preserves structural contrast and rank order among high-capacity edges.

> ***Remark* C.1** (Failure Modes and Graceful Degradation)**.** *Without OT calibration, the empirical distribution of raw kernel weights lacks natural scale normalization: maximum edge weights can grow unboundedly depending on the choice of divergence d and kernel κ, while sparse regions may be ill-conditioned. This directly destabilizes the GNN weight operator's spectrum, causing stability and discriminability bounds to behave unpredictably—particularly through deep layers where stochastic gradient variance compounds. The localized bounds $m(\mathcal{C})$ and $M(\mathcal{C})$ address this by providing regime-specific control: $M(\mathcal{C})$ ensures noise fluctuations in the dense head are damped, while $m(\mathcal{C})$ guarantees that meaningful stylistic contrasts in the sparse tail are preserved. When preprocessing or data characteristics yield overlapping supports with no clear structural partition between noise and signal, the localized framework cannot selectively partition regimes. In this case, OT calibration gracefully falls back to global operator norm regularization, acting as a mass-preserving smoother that bounds overall variance while relying on the task-driven feature representations in H to compensate for structural ambiguity.*

D Further Experimental Details

D.1 Experimental Details

In this subsection, we provide full experimental details. Code is available at `https://anonymous.4open.science/r/LLM-fixed-GNN-trained-Authorship-Attribution-ECF8`.

D.1.1 Datasets and Preprocessing

Authorship corpus. The primary corpus consists of 154 plays by six Early Modern English playwrights: Shakespeare, Fletcher, Jonson, Middleton, Chapman, and Marlowe (Segarra et al., 2015). Each play is segmented into contiguous, non-overlapping chunks of approximately $5{,}000$ words, with trailing remainders shorter than one chunk discarded. This yields 773 nodes with a majority-class baseline of 0.308.

Content masking. Each chunk is processed using part-of-speech and named-entity tagging. Content words are replaced by their POS or entity-type tags, while function words and punctuation are retained. The same masked text is used for both WAN construction and semantic feature extraction.

Auxiliary corpora. For author location, we use 12 novels by eleven nineteenth-century novelists: six American (Hawthorne, Melville, Twain, Alcott, Cooper, and Stowe) and five English (Austen, Dickens, Eliot, Trollope, and Hardy), obtained from Project Gutenberg. The opening chunk of each book is discarded to remove front matter, yielding 345 nodes with a majority-class baseline of 0.530. For time-period classification, we use 96 English plays from three periods: Early Modern, Victorian, and Modern. The Victorian group consists of works by Wilde, Shaw, Gilbert, Pinero, Jones, and Boucicault, and the Modern group of works by O'Neill, Synge, Gregory, Granville-Barker, Glaspell, Galsworthy, Barrie, and Milne. This yields 608 nodes with a majority-class baseline of 0.387. All

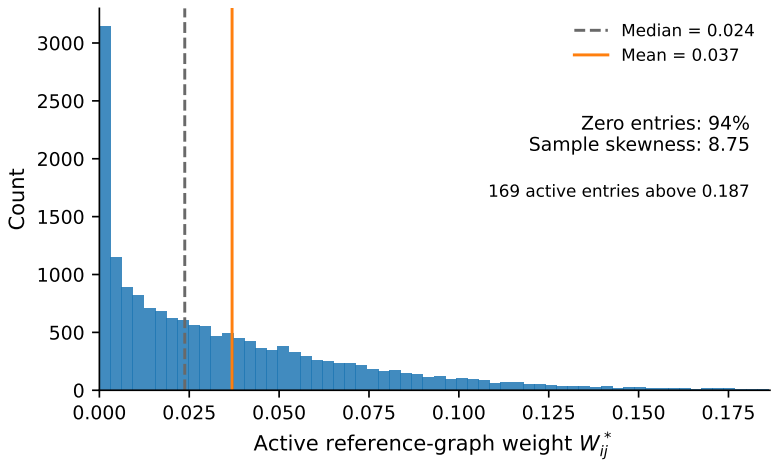

Figure 3: Empirical distribution of the active entries of the learned reference graph W^*. The graph contains 94% zero entries, and its active weights have sample skewness 8.75. The histogram is truncated at the 99th percentile for readability.

three period groups are restricted to plays to control for genre. All new data is obtained from Project Gutenberg.[2]

D.1.2 Word Adjacency Networks

WANs are constructed using the fixed lexicon of 30 function words from Segarra et al. (2015). For each chunk, every ordered pair of function words (f_i, f_j) occurring within $D = 10$ tokens in the same sentence contributes α^{d-1} to the corresponding transition weight, where d is the token separation and $\alpha = 0.75$. Sentences are delimited by `.`, `?`, `!`, and `;`. Row-normalizing the resulting counts gives the transition matrix $\hat{P} \in \mathbb{R}^{30\times 30}$. Rows with no observed transitions are replaced by the uniform distribution $1/|F|$. The stationary distribution $\hat{\pi}$ is computed by power iteration, with the Cesàro average used as a fallback for non-ergodic chains, and is normalized to sum to one.

D.1.3 Graph Construction

Pairwise divergences are computed between node stationary distributions, and each node is connected to its $k = 3$ nearest neighbors under the selected divergence, excluding itself. Nearest neighbors are obtained from the full pairwise distance matrix. Distances are mapped to edge weights using the regularized inverse kernel $\kappa(d) = (1 + d)^{-1}$. The directed k-NN graph is symmetrized by averaging w_{ij} and w_{ji}, with a missing direction treated as zero. The raw and OT-calibrated GCNs use the same symmetrized edge set and differ only in their edge weights.

D.1.4 Optimal-Transport Calibration

The reference graph W^* is obtained by solving the Kalofolias objective in Section 6.1 on the frozen semantic node features, with $\alpha = \beta = 1$. We use L-BFGS-B with at most $4{,}000$ iterations and tolerance 10^{-9}. Features are standardized by dimension and the squared-distance matrix Z is normalized by its median before optimization. Calibrated weights are obtained by one-dimensional quantile matching of the active WAN weights onto the active entries of W^*. Figure 3 shows the empirical distribution of the learned reference-graph weights. The graph contains 94% zero entries, while its active weights are strongly right-skewed, with sample skewness 8.75.

[2] https://www.gutenberg.org

D.1.5 WASSERSTEIN GROUND METRIC

For W_1 and W_2 in Table 2, the ground metric between function words is constructed from the per-text WAN transition matrices. We sum these matrices across the corpus and symmetrize the result to obtain a reference matrix S. Pairs with $S_{xy} > 10^{-12}$ are assigned edge cost $1/S_{xy}$, while all other pairs are assigned infinite cost. The ground distance $d(x, y)$ is the undirected shortest-path distance in this weighted graph. Finite distances are normalized by their maximum value, while disconnected pairs are assigned a normalized distance of 1. Wasserstein distances are computed using POT's `emd2` solver with cost matrix d^k.

D.1.6 MODEL ARCHITECTURES

Node features. Each chunk is split into non-overlapping windows of 256 tokens and encoded using frozen `sentence-transformers/all-MiniLM-L6-v2`. Token embeddings within each window are mean-pooled to obtain a 384-dimensional representation. Window representations are summarized by concatenating their elementwise mean and maximum, producing a 768-dimensional descriptor for each node. The encoder remains frozen throughout training.

Classifiers. Models using semantic features share a two-layer window aggregator ($768 \to 384 \to 384$) with GELU activation, trained jointly with the classifier. The MLP baseline applies a single linear layer to the aggregated features, while the GCN variants use a single `GCNConv` layer. Both map the 384-dimensional representation directly to the number of classes and use LeakyReLU with negative slope 0.01. The WAN-only baseline applies a single linear layer to the standardized stationary distribution $\hat{\pi}$ and does not use semantic features.

D.1.7 TRAINING

All models are implemented in PyTorch and PyTorch Geometric and trained with Adam using a learning rate of 10^{-3}, weight decay 5×10^{-4}, and cross-entropy loss. We use a $60/20/20$ node-level train/validation/test split, redrawn for each seed and shared across compared models within each seed. For authorship attribution, models are trained for up to 200 epochs with early stopping on validation loss and patience 20, and evaluated at the best validation checkpoint. For location and time-period classification, checkpoint selection and early stopping use validation accuracy, with patience 100 and a maximum of $1{,}000$ epochs. Results are reported as mean $\pm$ standard deviation over 20 random splits for authorship attribution and 10 random splits for each auxiliary task. Statistical comparisons use paired t-tests over matched splits.

D.1.8 ROBUSTNESS TO WAN PERTURBATIONS

Each transition matrix is perturbed as $\tilde{P} = (1 - \varepsilon)\hat{P} + \varepsilon R$, where rows of R are sampled from a symmetric Dirichlet$(1, \ldots, 1)$ distribution. We evaluate $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ with 10 noise seeds per level. For each divergence, perturbation level, and seed, the graph is recomputed and the GCN retrained with both raw and OT-calibrated weights; semantic features remain fixed. We evaluate TV, Hellinger, Jensen–Shannon, Fisher–Rao, KL, and W_1, with accuracy slopes estimated by least-squares regression on ε.

D.2 DIVERGENCE GEOMETRY

To validate Theorem 4.1, we transfer a fixed mass ε between function words ($\|v\|_1 = 2\varepsilon$) in the 773 stationary distributions. Growth exponents are estimated by log–log regression over $\varepsilon \in [10^{-4}, 10^{-2}]$. For the rare/common comparison ($\varepsilon = 10^{-3}$), pairs are matched on ground-metric distance so only baseline probability differs.

Table 2: Empirical validation of Theorem 4.1. Parentheses denote theoretical predictions.

	TV	W_1	Hell.	JS	FR	W_2
Growth exponent	1.000 (1)	0.993 (1)	0.946 (1)	0.947 (1)	0.946 (1)	0.495 (1/2)
Rare/common	1.00 (1)	1.00 (1)	2.59 (>1)	2.59 (>1)	2.59 (>1)	0.98 (1)

Results (Table 2) confirm the predicted linear scaling of TV, W_1, Hellinger, Jensen–Shannon, and Fisher–Rao, and the $\varepsilon^{1/2}$ scaling of W_2. The rare/common ratio validates the Fisher-information dependence: Hellinger, Jensen–Shannon, and Fisher–Rao are $\sim 2.6\times$ more sensitive to rare-word perturbations, while TV and Wasserstein remain unchanged.