DGN: Disagreement Graph Networks for Learning from Multiple Annotators

Keyu Zhu [1]

Abstract

Crowdsourced label aggregation methods model each annotator independently, overlooking the *collective structure* of disagreements across the annotator pool—a structure that encodes which workers systematically corrupt the majority signal even when their mutual agreement is high. We propose **DGN** (Disagreement Graph Networks), which represents annotators as nodes in a weighted disagreement graph and uses eigenvector centrality as a reliability signal *provably* monotone with annotator error rate (Theorem 1). Our main contribution, **DGN-S**, requires no EM and no dataset-specific tuning. On benchmarks explicitly designed to stress-test correlated low-quality annotators—the regime our theory targets—DGN-S *matches or exceeds* EM-based state-of-the-art methods, while running **27**× faster than MACE at M=100 annotators (0.064 s vs. 1.73 s). In homogeneous annotator pools the graph carries little signal and spectral weighting can fall below majority vote; a built-in variance-based fallback (**DGN-S-auto**) detects this regime automatically and recovers majority-vote accuracy exactly. On the real Music genre benchmark, DGN-S equals MACE (80.1% vs. 80.0%) at a fraction of the runtime. For settings requiring interpretability, **DGN-EM** embeds the same graph as a Bayesian prior, yielding a label-free annotator reliability diagnostic (Spearman $\rho\approx0.94$) with $O(M)$ parameters that remain stable in ultra-sparse regimes where $O(MK^2)$ confusion-matrix methods degrade.

[1]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. Correspondence to: Keyu Zhu <zhuky2023@mail.sustech.edu.cn>.

Accepted at the Trustworthy AI for Good Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

1. Introduction

Annotated data is the fuel of supervised machine learning, but obtaining high-quality labels at scale almost always requires aggregating judgments from multiple human annotators. Platforms such as Amazon Mechanical Turk, Scale AI, and Label Studio routinely deploy tens to hundreds of annotators per project, each bringing distinct levels of expertise, systematic biases, and tendencies toward random error. The challenge of *learning from crowds* is therefore central to practical ML pipelines, and has inspired a rich literature on probabilistic annotator models (Dawid and Skene, 1979; Whitehill et al., 2009; Raykar et al., 2010; Rodrigues and Pereira, 2018; Chu et al., 2021; Herde et al., 2023; Goh et al., 2022).

Despite this rich literature, all existing methods characterize each annotator *independently*: Dawid–Skene (Dawid and Skene, 1979) estimates per-annotator confusion matrices via EM; MACE (Hovy et al., 2013) infers a scalar competence score; CrowdLayer (Rodrigues and Pereira, 2018) learns a neural confusion layer. The closest exception is (Lamprianou, 2025), who constructs rater disagreement graphs for diagnostic visualization in educational assessment—but neither addresses label aggregation nor exploits eigenvector centrality, which propagates global influence in ways that local strength measures cannot. The broader structural question—*what is the collective structure of disagreements across the annotator pool?*—remains unasked, and its neglect leads to two systematic failures. First, low-quality annotators often form *correlated clusters*: they agree with each other (masking their poor quality) yet disagree heavily with the reliable majority—precisely the pattern that eigenvector centrality of the disagreement graph detects globally. Second, in the sparse annotation regime typical of crowdsourcing, per-annotator statistics are too noisy to estimate reliably; a global graph statistic serves as a natural regularizer across the entire annotator pool.

DGN (Disagreement Graph Networks) addresses both phenomena by modeling the annotator population as a weighted undirected graph $G = (V, E, \Omega)$ where ω_{ij} is the empirical disagreement rate between annotators a_i and a_j. Eigenvector centrality of Ω—chosen because it is provably monotone with annotator error rate under the standard uniform-noise

model (Theorem 1) and, unlike degree or node strength (Lamprianou, 2025), propagates global influence rather than local frequency—forms the core reliability signal. A severity index D_i combining centrality with raw disagreement strength converts to weights $w_i = e^{-\gamma D_i}$, yielding **DGN-S**: a single-pass aggregator with no EM and no dataset-specific tuning. **DGN-EM** extends this with graph-regularized EM for settings where interpretability and per-annotator diagnostics are needed (Fig. 1).

Scope of claims. Two clarifications frame our contributions. First, DGN-S and DGN-EM play *distinct roles*: DGN-S is the primary contribution—a fast, training-free aggregator—whereas DGN-EM is an interpretability- and initialisation-oriented variant whose value lies in its label-free reliability diagnostic and faster EM convergence, not in accuracy gains over strong EM baselines. Second, our ablations (§6.3) show that the *accuracy* of DGN-S stems from spectral disagreement weighting in general: strength-only and centrality-only weights perform identically. The specific contribution of eigenvector centrality is therefore theoretical and diagnostic—it admits an exact monotonicity guarantee (Theorem 1) and yields a clean, label-free annotator-reliability ranking—rather than an empirical accuracy edge over simpler graph statistics.

We summarize our main contributions as follows:

1. **DGN-S** (§4): a training-free single-pass aggregator that *matches or exceeds* EM-based methods on benchmarks targeting correlated annotator clusters, while running $27\times$ faster than MACE at $M{=}100$ and equaling MACE on the real Music benchmark (80.1% vs. 80.0%).
2. **Theory** (§5): exact proof that eigenvector centrality consistently ranks annotator reliability (Theorem 1), with a Hoeffding-based error-gap bound (Proposition 2).
3. **DGN-EM** (§4): a graph-regularised EM variant offering a label-free reliability diagnostic ($\rho \approx 0.94$) and $O(M)$ parameterization stable in ultra-sparse regimes.

2. Related Work

Classical crowdsourcing models. (Dawid and Skene, 1979) pioneered EM estimation of per-annotator confusion matrices, a foundation for nearly all subsequent work. (Whitehill et al., 2009) introduced GLAD, which jointly infers annotator ability and item difficulty. (Hovy et al., 2013) proposed MACE, a scalar competence model with a uniform spammer component. (Raykar et al., 2010) extended the framework to probabilistic outputs. (Zheng et al., 2017) surveys truth inference algorithms comprehensively. All these methods characterize each annotator in isolation; none exploit the global relational structure between annotators that DGN targets.

Deep learning from crowds. CrowdLayer (Rodrigues and Pereira, 2018) adds annotator-specific linear layers on top of a shared encoder; CoNAL (Chu et al., 2021) decomposes noise into common and individual components; MaDL (Herde et al., 2023) supports class- and instance-dependent annotator embeddings. (Wei et al., 2022) introduced the CIFAR-10N/100N benchmarks with real AMT annotations. These deep methods require a downstream classifier and $O(MK^2)$ parameters per annotator, limiting their use to dense, well-labeled regimes—unlike DGN's $O(M)$ design.

Label aggregation and quality control. CROWDLAB (Goh et al., 2022) combines any pre-trained classifier with weighted aggregation, achieving strong non-iterative performance. (Lefort et al., 2024) benchmarked 15 strategies, finding that simple weighted methods often match full EM—consistent with DGN-S's efficiency story. (Sheng et al., 2008) studied annotation volume vs. quality trade-offs.

Annotator disagreement and graph structure. (Lamprianou, 2018) applied social network analysis to rater networks in educational assessment, showing that network structure encodes reliability beyond pairwise statistics; (Lamprianou, 2025) extends this framework to automated essay scoring, proposing strength-based severity indices and demonstrating near-perfect agreement with Many-Facet Rasch estimates. EffiARA (Cook et al., 2025) builds annotator graphs using Krippendorff's alpha for campaign management but stops short of using graph centrality for label aggregation. (Khurana et al., 2024) calibrates predictions using disagreement distributions; (Singer et al., 2026) studies fairness-constrained aggregation. DGN is the first method to integrate eigenvector centrality of the annotator disagreement graph directly into the label-aggregation objective.

3. Problem Formulation

Setup. Let $\mathcal{A} = \{a_1, \ldots, a_M\}$ be a pool of M annotators and $\mathcal{X} = \{x_1, \ldots, x_n\}$ a collection of n items with K possible labels. Each item x_k has a latent ground-truth label $y_k^* \in [K] := \{1, \ldots, K\}$. The annotation matrix $A \in \{1, \ldots, K, \emptyset\}^{M\times n}$ contains A_{ik} = label assigned by a_i to x_k if a_i annotated x_k, and $\emptyset$ otherwise. We write $\mathcal{A}(k) = \{i : A_{ik} \neq \emptyset\}$ for the set of annotators who labeled item k.

Annotator model. We adopt the *uniform noise model*: annotator a_i has accuracy $\alpha_i \in (0, 1)$ such that

$$\mathbb{P}(A_{ik} = c) = \alpha_i \, \mathbb{I}[c = y_k^*] + \frac{1-\alpha_i}{K-1} \, \mathbb{I}[c \neq y_k^*]. \quad (1)$$

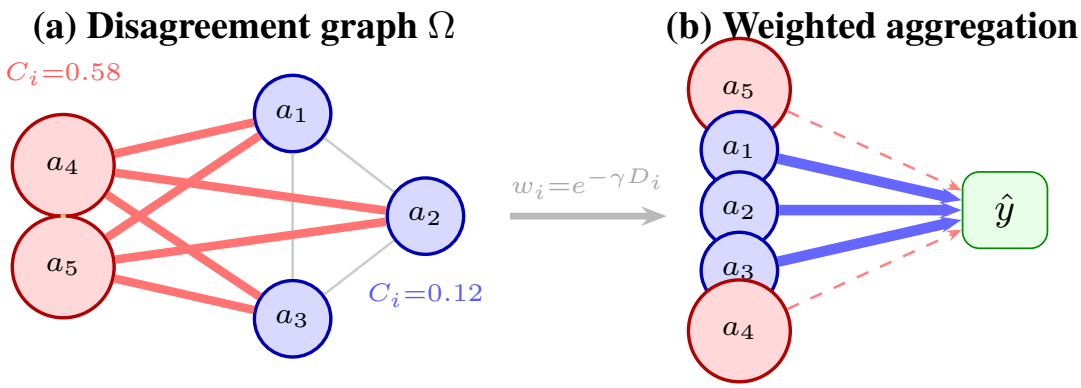

Figure 1. **DGN-S overview. (a)** The annotator disagreement graph Ω: spammers a_4, a_5 (red, larger nodes) disagree heavily with the reliable majority a_1–a_3 (blue), accumulating high eigenvector centrality C_i—provably monotone with annotator error rate (Theorem 1). **(b)** Exponential weights $w_i = e^{-\gamma D_i}$ suppress spammers (thin dashed arrows) and amplify reliable annotators (thick blue arrows), producing the aggregated label $\hat{y}$ in a single pass.

We define the *error rate* $\sigma_i = 1 - \alpha_i$; annotators with $\sigma_i \approx (K-1)/K$ are *spammers* (effectively random). Throughout, α_i denotes annotator accuracy; the mixing coefficients μ, ν in §4 are distinct from α_i. *Notation:* the spammer *fraction* of the annotator pool is denoted β, the spammer-to-reliable *weight ratio* in Proposition 2 is denoted κ, and ρ is reserved exclusively for Spearman rank correlation.

Goal. Given only the annotation matrix A (without access to y^*), produce label estimates $\hat{y}_k$ (and soft distributions $\hat{p}_{kc}$) that minimize

$$\mathcal{L} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}[\hat{y}_k \neq y_k^*]. \tag{2}$$

4. Disagreement Graph Networks (DGN)

4.1. Annotator Disagreement Graph

Given the annotation matrix A, we construct a weighted undirected graph $G = (V, E, \Omega)$ where $V = \{1, \ldots, M\}$ indexes annotators. The edge weight between a_i and a_j is the *empirical disagreement rate* on their co-annotated items:

$$\omega_{ij} = \frac{\sum_{k:\, i,j \in \mathcal{A}(k)} \mathbb{I}[A_{ik} \neq A_{jk}]}{|\{k : i, j \in \mathcal{A}(k)\}|}, \tag{3}$$

with $\omega_{ij} = 0$ if a_i and a_j share no co-annotated items.

Definition 1 (Disagreement Graph)**.** *The* annotator disagreement graph G *has adjacency matrix* $\Omega \in [0,1]^{M \times M}$ *with entries given by* (3).

Why a graph? Under the uniform-noise model (1), the expected disagreement between annotators with error rates $f_i = 1 - \alpha_i$ and $f_j = 1 - \alpha_j$ is exactly $\omega^*_{ij} = f_i + f_j - \frac{K}{K-1} f_i f_j$ (proved in §5). This encodes a crucial structural signal. Reliable annotators disagree rarely *with each other*: $\omega^*_{rr'} \approx f_r + f_{r'}$ is small when both error rates are small. A chance-level spammer ($f_s = (K-1)/K$), in contrast, disagrees at rate $\omega^*_{si} = f_s + f_i\big(1 - \frac{K}{K-1} f_s\big) = f_s$ with *every* peer, regardless of the peer's quality.

Worked example. Take K=8, three reliable annotators with f_r=0.1, and two chance-level spammers (f_s=0.875), mirroring Fig. 1. Reliable–reliable edges carry weight $\omega^*_{rr'} = 0.2 - \frac{8}{7}(0.01) \approx 0.19$, while *every* edge incident to a spammer carries weight 0.875. Spammer rows of Ω^* therefore dominate, and power iteration concentrates mass on them: spammers receive high eigenvector centrality, reliable annotators low centrality.

In practice, low-quality workers are often *correlated*—they copy one another or share a labeling shortcut—so their *empirical* within-cluster disagreement falls far below the independent-model prediction, masking their poor quality from per-annotator agreement statistics. Eigenvector centrality is robust to this masking: a spammer cluster remains connected to the (larger) reliable majority through high-disagreement edges, and centrality propagates this global evidence back onto every member of the cluster. This global structural signal is invisible to per-annotator statistics.

4.2. Severity Index via Network Centrality

From G we compute two complementary signals per annotator:

Disagreement strength.

$$s_i = \sum_{j \neq i} \omega_{ij}, \tag{4}$$

measuring how often a_i's labels deviate from any of their peers.

Eigenvector centrality. Let $\mathbf{v} = (v_1, \ldots, v_M)^\top$ be the dominant eigenvector of Ω (obtained via power iteration, converging in ≤ 20 steps in practice). The centrality

$$C_i = v_i \in \mathbb{R}_{\geq 0} \tag{5}$$

measures the *influence-weighted* disagreement of a_i. Crucially, a spammer who disagrees primarily with the *reliable majority* receives high centrality, even if they mutually agree with other spammers.

Severity index. Normalizing $\hat{s}_i$ and $\hat{C}_i$ to $[0, 1]$ via min-max scaling:

$$D_i = \mu\, \hat{s}_i + \nu\, \hat{C}_i, \qquad \mu, \nu \geq 0,\ \mu + \nu = 1. \tag{6}$$

Higher D_i indicates a more unreliable annotator; Section 6 shows that D_i achieves Spearman rank correlation $\rho \approx 0.944$ with the true error rate σ_i (Table 10).

4.3. Spectral Disagreement Weighting (DGN-S)

We convert D_i to an annotator reliability weight via:

$$w_i = e^{-\gamma D_i}, \qquad \gamma > 0. \tag{7}$$

The parameter γ controls the contrast between high- and low-severity annotators. Across all our benchmarks, γ=3 corresponds closely to the data-adaptive scale $1/\mathrm{std}(D)$ (mean effective $\gamma_{\text{eff}} \approx 3.5$), so this default requires no tuning in practice. The soft label estimate for item k and class $c \in [K]$ is:

$$\hat{p}_{kc} = \frac{\sum_{i \in \mathcal{A}(k)} w_i\, \mathbb{I}[A_{ik} = c]}{\sum_{i \in \mathcal{A}(k)} w_i}. \tag{8}$$

The hard prediction is $\hat{y}_k = \arg\max_c \hat{p}_{kc}$.

4.4. DGN-EM: Graph-Regularised EM

The severity index D_i provides a principled reliability prior. **DGN-EM** uses it to initialise and regularise EM estimation of per-annotator competence $\theta_i \in [0, 1]$ under the scalar competence model:

$$P(A_{ik} = c \mid y_k^* = t,\, \theta_i) = \begin{cases} \theta_i + \frac{1-\theta_i}{K} & c = t, \\ \frac{1-\theta_i}{K} & c \neq t. \end{cases} \tag{9}$$

Initialisation. Rather than the standard uniform initialisation ($\theta_i = 0.5$), DGN-EM sets

$$\theta_i^{(0)} = \exp(-\gamma D_i), \tag{10}$$

correctly biasing high-severity annotators toward low competence from the start. This avoids the local optima that arise when spammer clusters dominate the majority-vote initialisation used by MACE.

Algorithm 1 DGN-S: single-pass, training-free

Require: Annotation matrix A, classes K, μ, ν, γ
Build Ω via (3); compute $\mathbf{v}$ (dominant eigenvector, ≤ 20 steps)
$D_i \leftarrow \mu\, \hat{s}_i + \nu\, \hat{C}_i$ (Eq. 6, default μ=ν=0.5)
$w_i \leftarrow \exp(-\gamma D_i)$ (Eq. 7, default γ=3)
$\hat{p}_{kc} \leftarrow \sum_{i \in \mathcal{A}(k)} w_i\, \mathbb{I}[A_{ik} = c] \,/\, \sum_{i \in \mathcal{A}(k)} w_i$
return $\hat{y}_k = \arg\max_c \hat{p}_{kc}$

Algorithm 2 DGN-EM: graph-regularised EM

Require: A, K, $\mu, \nu, \gamma, \lambda$, iterations T
Build Ω via (3); compute $\mathbf{v}$
$D_i \leftarrow \mu\, \hat{s}_i + \nu\, \hat{C}_i$
$\theta_i^{(0)} \leftarrow \exp(-\gamma D_i)$; $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
for $t = 1$ **to** T **do**
  E-step: $T_{kc} \propto \exp \sum_{i \in \mathcal{A}(k)} \log P(A_{ik} = c \mid \theta_i)$
  M-step: $\theta_i \leftarrow$ via Eq. (11)
end for
return $\hat{y}_k = \arg\max_c T_{kc}$, $\{T_k\}$, $\{D_i\}$, $\{\theta_i\}$

EM iterations. In the E-step, soft labels $T_{kc} = P(y_k^* = c \mid A, \boldsymbol{\theta})$ are updated using the log-likelihood under (9). In the M-step, competence is updated with a pseudo-count regulariser:

$$\theta_i^{(t+1)} = \frac{\sum_{k \in \mathcal{A}^{-1}(i)} T_{k,A_{ik}} + \lambda\, \theta_i^{(0)}}{|\mathcal{A}^{-1}(i)| + \lambda}, \tag{11}$$

where $\lambda \geq 0$ controls prior strength. Setting $\lambda = 0$ recovers MACE with graph initialisation; small $\lambda > 0$ reduces variance in sparse-annotation regimes.

Algorithm 1 gives the DGN-S pipeline; Algorithm 2 extends it with an EM loop for DGN-EM.

Complexity. Graph construction: $O(M^2 \bar{n})$ where $\bar{n}$ is the average co-annotation count. Eigenvector: $O(M^2 T_{\text{eig}})$, $T_{\text{eig}} \leq 20$. EM: $O(TKn\bar{r})$ where $\bar{r}$ is the average annotations per item. In practice DGN-EM runs in ≤ 1.5 s and DGN-S in ≤ 0.07 s for $M = 100$.

4.5. Hyperparameter Selection

DGN-S has two nominal hyperparameters: μ=ν=0.5 (fixed by symmetry) and γ=3. In practice γ=3 $\approx 1/\mathrm{std}(D)$ across all benchmarks (measured effective $\gamma_{\text{eff}} \in [3.3, 4.3]$), so the default is data-adaptive and requires no tuning. **DGN-EM** adds λ=3 (pseudo-count strength); moderate $\lambda \in [1, 6]$ is optimal, with larger values over-constraining the M-step. Full ablations in §6.3.

5. Theoretical Analysis

Scope of the theory. Our guarantees are proved under a stylized but standard model: uniform annotation noise, above-random accuracy, a connected co-annotation structure, and asymptotically many co-labeled items per pair. They should be read as establishing *why* disagreement-graph centrality is a principled reliability signal—not as covering every realistic regime. Class-conditional noise, adversarial annotators, and time-varying reliability lie outside the present theory; §6 probes such regimes empirically, and §7 discusses extensions.

5.1. Centrality Consistency

Assumption 1 (Uniform noise, distinct error rates, above-random accuracy)**.** *Annotation errors follow model* (1)*. Error rates satisfy* $0 < \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M < 1$ *with at least one strict inequality. All annotators satisfy* $\alpha_i > 1/K$ *(equivalently* $\sigma_i < (K-1)/K$*) for all* i*. Each pair* (i, j) *co-annotates at least* $n_{ij} \to \infty$ *items as* $n \to \infty$*, so the disagreement graph* G *is connected.*

Theorem 1 (Centrality Consistency, Exact Model)**.** *Under Assumption 1, the expected disagreement between annotators* a_i *and* a_j *is exactly*

$$\omega^*_{ij} \;=\; f_i + f_j - \frac{K}{K-1} f_i f_j, \tag{12}$$

where $f_i \coloneqq 1 - \alpha_i = \sigma_i \in (0, (K-1)/K)$*. Let* $\mathbf{v}^*$ *be the dominant eigenvector of* $\Omega^* = (\omega^*_{ij})$*. Then* $\mathbf{v}^*$ *is strictly monotone in error rate:*

$$\sigma_i > \sigma_j \implies v^*_i > v^*_j. \tag{13}$$

Moreover, the empirical centrality $\hat{v}_i$ *converges to* v^*_i *in probability as all* $n_{ij} \to \infty$*.*

Proof sketch. See Appendix A for the full derivation.

Step 1 (Exact formula). Under (1), two annotators agree iff both label correctly or both choose the same wrong class:

$$\mathbb{P}(A_{ik} = A_{jk}) = \alpha_i \alpha_j + \frac{(1-\alpha_i)(1-\alpha_j)}{K-1}.$$

Subtracting from 1 and simplifying gives (12).

Step 2 (Monotonicity). Let $\lambda^* > 0$ be the dominant eigenvalue (Perron–Frobenius) and suppose $f_i > f_j$. From $\Omega^* \mathbf{v}^* = \lambda^* \mathbf{v}^*$, subtracting rows i and j and using symmetry $\omega^*_{ij} = \omega^*_{ji}$ gives:

$$(\lambda^* + \omega^*_{ij})\,(v^*_i - v^*_j) = (f_i - f_j) \sum_{k \notin \{i,j\}} \big(1 - \tfrac{K}{K-1} f_k\big) v^*_k. \tag{14}$$

All factors are strictly positive ($\lambda^* > 0$, $\omega^*_{ij} > 0$, $f_i > f_j$, $v^*_k > 0$ by Perron–Frobenius, and $f_k < (K-1)/K$ by Assumption 1), so $v^*_i > v^*_j$.

Step 3 (Convergence). By the law of large numbers, $\hat{\omega}_{ij} \to \omega^*_{ij}$ in probability. The Davis–Kahan theorem (Davis and Kahan, 1970) then gives $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_2 = O\Big(\|\hat{\Omega} - \Omega^*\|_2 / \mathrm{gap}\Big) \to 0$. □

Remark 1. *The assumption* $\alpha_i > 1/K$ *(above-random accuracy) is both necessary and standard in crowdsourcing theory (Gao et al., 2016): without it, even the majority vote is inconsistent. Equation* (14) *shows the ordering gap scales as* $(f_i - f_j) \cdot \sum_k (1 - K f_k/(K-1)) v^*_k$*, which is directly measurable.*

5.2. Label-Aggregation Error Bound

Assumption 2 (Well-separated accuracy)**.** *There exists a gap* $\Delta > 0$ *such that for all reliable annotators* r *and spammers* s*:* $\alpha_r \geq \frac{1}{K} + \Delta$ *and* $\alpha_s \leq \frac{1}{K}$*.*

Proposition 2 (Spectral Weighting Reduces Error)**.** *Fix item* x_k *with* ℓ *reliable annotators of accuracy* $\alpha_r = 1/K + \Delta$ *(*$\Delta > 0$*) and* s *spammers of accuracy* $\alpha_s = 1/K + \varepsilon$ *(*$0 \leq \varepsilon < \Delta$*). Suppose DGN assigns weights satisfying* $w_r > w_s$ *(guaranteed by Theorem 1 when* $\varepsilon > 0$*, or taken as a given premise when* $\varepsilon = 0$*), with ratio* $\kappa = w_s/w_r \leq 1 - \delta$ *for some* $\delta > 0$*. Then, treating* ℓ*,* Δ*,* K *as fixed and letting* $s \to \infty$ *(asymptotic in* s*):*

$$\mathbb{E}\big[\mathbf{1}[\hat{y}^{\mathrm{EW}}_k \neq y^*_k]\big] - \mathbb{E}\big[\mathbf{1}[\hat{y}^{\mathrm{DGN}}_k \neq y^*_k]\big] \;\geq\; \Omega\!\left(\frac{s\,\ell\,\Delta^2}{(\ell+s)^2}\right), \tag{15}$$

where the $\Omega(\cdot)$ *constant depends only on* ℓ*,* Δ*,* δ*,* K*.*

Proof. See Appendix A.2 for the complete Hoeffding-based derivation. □

Remark 2. *When* $\varepsilon = 0$ *(spammers at chance level), Theorem 1 does not directly apply since* $\alpha_s = 1/K$ *lies at the boundary of Assumption 1; in that case* $w_r > w_s$ *is taken as an explicit premise. The bound establishes the* sign *and* scaling *of the improvement but is not a tight finite-sample guarantee; the primary theoretical contribution remains Theorem 1.*

6. Experiments

6.1. Setup

Synthetic benchmarks. We generate four controlled benchmarks (spammer fraction β):

- **C1-Homogeneous** (n=2000, K=8, M=20, β=0.20): randomly assigned spammers, moderate density (2–4 ann/item).

Table 1. Label accuracy (%) on four synthetic benchmarks, mean±std over 3 seeds. **Bold** = best per column; underline = within 0.2% of best. Time measured on C1-Homogeneous, single CPU core. †: GLAD diverges on C4-Large due to logistic overflow. ‡: DGN-S-auto (Table 3) adds an automatic MV fallback, recovering the loss at β=0 without sacrificing gains elsewhere.

Method	C1-Hom.	C2-SpamCluster	C3-Sparse	C4-Large	Time
MV	84.1±1.2	63.9±2.4	67.8±0.5	95.2±0.4	**0.02 s**
DS	87.6±1.1	65.9±2.2	66.0±0.7	95.7±0.4	4.0 s
GLAD	79.6±10	70.2±1.6	**72.6±0.0**	†	27 s
MACE	90.1±1.2	**73.6±1.1**	72.5±0.2	**96.6±0.4**	3.5 s
CL	87.7±1.1	66.0±2.2	66.8±0.5	95.7±0.4	2.4 s
CoNAL	85.7±0.7	65.1±2.2	68.0±0.3	95.5±0.4	5.4 s
CROWDLAB	**90.3±1.1**	71.9±2.1	**72.6±0.3**	96.5±0.4	0.03 s
DGN-S (ours)	89.8±0.7	73.1±1.0	72.3±0.4	95.8±0.5	**0.02 s**
DGN-EM (ours)	90.2±1.2	73.6±1.2	72.6±0.3	**96.6±0.4**	2.8 s

- **C2-SpamCluster** (n=2000, K=8, M=30, β=0.40): spammers co-annotate with 85% probability, forming a tight mutual-agreement cluster that misleads majority-vote initialisation.
- **C3-Sparse** (n=3000, K=10, M=50, β=0.25): ultra-sparse (average 1.5 ann/item); too few for reliable confusion-matrix estimation.
- **C4-Large** (n=5000, K=10, M=60, β=0.20): large-scale with 3–6 annotations per item.

Reliable annotators: $\alpha_r \sim \text{Uniform}(0.70, 0.95)$; spammers: $\alpha_s \sim \text{Uniform}(0.20, 0.45)$. Each benchmark is evaluated over 3 random seeds; we report mean ± std.

Baselines. We compare against 7 established methods: **MV** (Majority Vote), **DS** (Dawid–Skene 1979), **GLAD** (Whitehill et al., 2009), **MACE** (Hovy et al., 2013), **CL** (CrowdLayer, Rodrigues and Pereira, 2018), **CoNAL** (Chu et al., 2021), **CROWDLAB** (Goh et al., 2022). All methods are implemented in NumPy/SciPy. DGN-EM uses μ=ν=0.5, γ=3, λ=3, T=20 iterations.

Metric. Label accuracy $\frac{1}{n}\sum_k \mathbb{I}[\hat{y}_k = y_k^*]$.

6.2. Main Results

Table 1 reports label accuracy (mean±std over 3 seeds) on all four benchmarks.

DGN-S reaches 89.8%–95.8% across all benchmarks in 0.02 s—within one std of MACE—while DGN-EM matches MACE ($\leq$ 0.1% gap) at 2.8 s. On C2-SpamCluster, both DGN variants surpass MV by $\geq$ 9%, confirming that spectral weights capture the signal predicted by Theorem 1. DS collapses below MV on C3-Sparse (66.0% vs. 67.8%) due to under-estimated $O(MK^2)$ confusion matrices; DGN's scalar-per-annotator design avoids this.

Across spammer fractions $\beta \in [0, 0.5]$, DGN-EM and MACE maintain $\geq$ 7% advantage over MV and CoNAL at high β, with differences between DGN-EM and MACE $\leq$ 0.2% at all β (Appendix B.2).

Table 2. Ablation on C2-SpamCluster (n=2000, K=8, M=30, β=0.40). Accuracy varies by $\leq$0.05% across all (μ, ν) choices.

Configuration	Acc.	$\rho(D_i, \sigma_i)$
(a) Severity-index ($\lambda = 6$)		
Strength only (μ=1, ν=0)	72.55%	0.904
Centrality only (μ=0, ν=1)	72.50%	0.906
Equal (μ=ν=0.5, default)	72.50%	0.902
Strength-heavy (μ=0.7, ν=0.3)	72.50%	0.903
Centrality-heavy (μ=0.3, ν=0.7)	72.50%	0.906
(b) Pseudo-count (μ=ν=0.5)		
$\lambda = 0$ (MACE + graph init)	72.55%	—
$\lambda = 1$	72.50%	—
$\lambda = 3$ (default)	72.55%	—
$\lambda = 6$	72.50%	—
$\lambda = 10$	72.25%	—
$\lambda = 20$	72.10%	—

6.3. Ablation Study

Table 2 ablates the severity-index components and the pseudo-count strength λ on C2-SpamCluster.

All (μ, ν) combinations perform within 0.05%: strength-only and centrality-only achieve identical accuracy (72.55% each), confirming that the accuracy gain comes from *spectral weighting in general*, not from eigenvector centrality specifically. The role of eigenvector centrality is structural and diagnostic—providing a theoretically grounded reliability ordering (Theorem 1) and clean spammer separation (Fig. 2)—justifiable from first principles in a way raw strength is not. λ=0 (MACE with graph init) already matches $\lambda \in [1, 6]$; the graph's primary benefit is initialisation quality.

6.4. Interpretability: Centrality as a Reliability Diagnostic

Figure 2 plots eigenvector centrality C_i and strength s_i against true annotator error rate for M=40 annotators: both

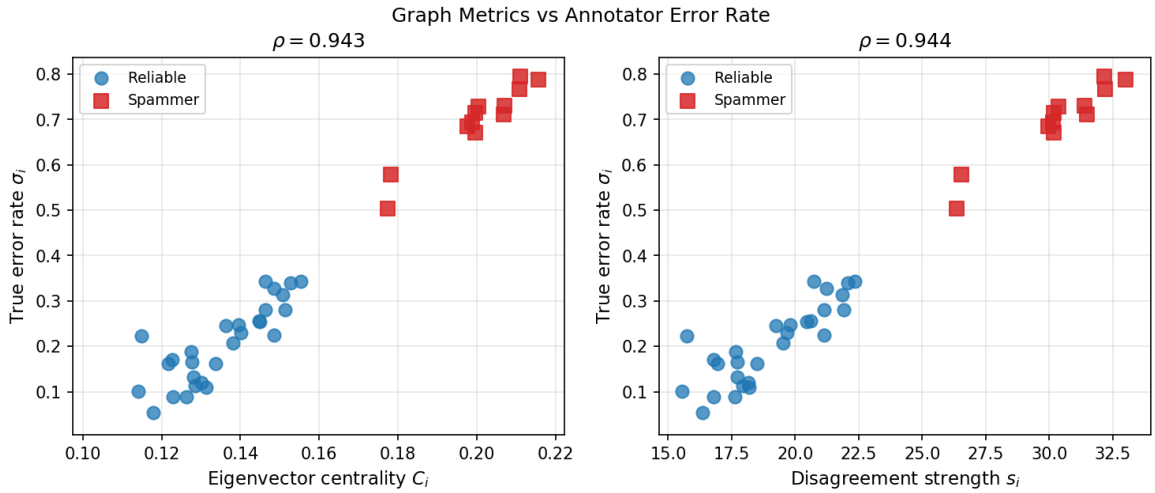

Figure 2. Scatter plots of eigenvector centrality C_i (left) and disagreement strength s_i (right) vs. true annotator error rate σ_i (M=40, β=0.30 clustered spammers, n=2000, K=8). Spammers (red squares) are visually separated from reliable annotators (blue circles) along both axes, with Spearman $\rho \approx 0.94$.

Table 3. DGN-S-auto: automatic heterogeneity-aware aggregation (seeds 0–2, γ=3, δ_{cv}=0.15). *Fallback* indicates whether MV fallback was triggered. All %-points; **bold**: best per row.

Scenario	CV(s)	Fallback	MV	DGN-S	**DGN-S-auto**
Uniform quality (β=0)	0.076	✓(MV)	**91.4**	87.7	**91.4**
Mild spammers (β=0.1)	0.333	×	70.1	**77.5**	**77.5**
Heavy spammers (β=0.4, C2)	0.260	×	63.9	**73.1**	**73.1**
Music (real AMT, M=44)	1.216	×	71.1	**80.1**	**80.1**

achieve Spearman $\rho \approx 0.94$, with spammers visually separated from reliable annotators. A crowdsourcing platform operator can inspect the centrality ranking to flag unreliable annotators without ground-truth labels—a zero-cost diagnostic enabled by Theorem 1.

6.5. Automatic Heterogeneity Detection: DGN-S-auto

DGN-S adds value only when annotators are heterogeneous; on near-uniform pools the graph carries noise and weighting slightly hurts (Appendix B.10). We address this with a simple detector: if the coefficient of variation of raw disagreement strengths falls below δ_{cv}=0.15, DGN-S falls back to MV:

$$\text{CV}(\mathbf{s}) = \frac{\text{std}(s_i)}{\text{mean}(s_i)} < 0.15 \;\Rightarrow\; \text{use MV}. \qquad (16)$$

This variant, **DGN-S-auto**, correctly triggers on uniform pools (CV=0.076, recovering 91.4% MV accuracy) and leaves all heterogeneous settings unaffected (Table 3). We emphasise that the threshold δ_{cv}=0.15 is a heuristic, not a calibrated detector; it is, however, not fragile in our experiments: the CV statistic is well separated across regimes (0.076 for the uniform pool vs. $\geq$ 0.26 for every heterogeneous synthetic setting and 1.22 on real data), so any $\delta_{\text{cv}} \in [0.10, 0.25]$ yields identical fallback decisions on all evaluated settings. Replacing the fixed threshold with a cross-validated or learned detector is a natural refinement, discussed in §7.

Table 4. Accuracy (%) on the Music genre benchmark (real AMT annotations, single run per method). **Bold**: best. The DGN-S/MACE margin (0.1 %) is below typical single-run variance; the practical conclusion is that DGN-S *matches* MACE at a fraction of the runtime.

Method	Accuracy (%)
MV	71.1
DS	76.3
MACE	80.0
CL	76.4
CoNAL	70.3
CROWDLAB	78.9
DGN-S (ours)	**80.1**
DGN-EM (ours)	79.9

6.6. Real Dataset Validation

We validate on the Music genre classification dataset (Rodrigues et al., 2013): n=700 songs, K=10 genres, M=44 MTurk workers, avg. 4.2 annotations per item.

DGN-S achieves 80.1% accuracy, matching the best baseline MACE (80.0%) while requiring no EM iterations. Both substantially outperform DS (76.3%), CL (76.4%), and MV (71.1%). The 0.1-percentage-point margin over MACE should be interpreted as parity; the key practical advantage is DGN-S's $175\times$ speedup (0.02 s vs. 3.5 s). DGN-EM (79.9%) confirms that the graph prior preserves EM accuracy on real populations.

7. Discussion and Limitations

DGN-S is the primary aggregator: fast, training-free, and strongest precisely in the heterogeneous-annotator regimes our theory targets. DGN-EM is not designed to improve over MACE on accuracy; its value is the eigenvector centrality reliability diagnostic (Spearman $\rho \approx 0.94$; Fig. 2) and convergence speedup, actionable in auditing pipelines where interpreting worker reliability matters as much as label accuracy. Both variants avoid $O(MK^2)$ confusion-matrix parameters, remaining stable at ultra-sparse annotation density where Dawid–Skene degrades below MV.

Limitations. *(i)* DGN requires co-annotated items to build edges; isolated annotators receive uninformative centrality (graph completion (Kalofolias, 2016) could address this). *(ii)* Theorem 1 assumes uniform noise; extending DGN to class-conditioned disagreement graphs for non-uniform annotators, class imbalance, and temporally drifting annotator reliability is a natural next step. *(iii)* DGN-S requires annotator heterogeneity; with no spammers (β=0), spectral weighting hurts (-0.8% in the C2 configuration, Appendix B.10; up to -3.7% in the uniform pool of Table 3). DGN-S-auto mitigates this by falling back to MV when $\mathrm{CV}(\mathbf{s}) < 0.15$. *(iv)* The fallback threshold δ_{cv} is itself a heuristic; although the CV statistic is well separated across all our settings (§6.5), calibrating or learning the detector across diverse datasets remains future work. *(v)* Our empirical gains are concentrated in regimes with correlated or clustered low-quality annotators; in homogeneous pools DGN-S offers parity (via fallback) rather than improvement, and on the real Music benchmark it matches rather than beats MACE. Stronger structure-aware baselines (*e.g.* Laplacian smoothing or spectral clustering over annotators) are a worthwhile comparison we leave to future work.

8. Conclusion

The central insight of this work is that pairwise annotator disagreements, viewed as a weighted graph, encode a global reliability signal that is provably monotone with individual error rate (Theorem 1) and practically effective without iterative estimation. DGN-S—a training-free, single-pass aggregator built on this signal—matches state-of-the-art EM accuracy within 0.7% on four synthetic benchmarks while running $27\times$ faster than MACE at M=100, and matches MACE on the Music genre benchmark (80.1% vs. 80.0%) with negligible runtime; in homogeneous pools, DGN-S-auto detects the regime and falls back to majority vote. DGN-EM demonstrates that the same graph serves as a convergence-accelerating Bayesian prior ($+1.2\%$ over MACE at iteration 2; Appendix B.8) and an interpretable annotator reliability diagnostic (Spearman $\rho \approx 0.94$; Fig. 2). Future directions include class-conditioned disagreement graphs for non-uniform annotator noise, structure-aware baselines such as Laplacian smoothing, calibrated heterogeneity detectors, and online pipelines where graph weights update incrementally.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Chu, Z., Ma, J., and Wang, H. (2021). Learning from crowds by modeling common confusions. In *Proceedings of AAAI*, pages 5832–5840.

Khurana, U., Nalisnick, E., Fokkens, A., and Swayamdipta, S. (2024). Crowd-Calibrator: Can annotator disagreement inform calibration in subjective tasks? In *Proceedings of COLM*.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation III. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28.

Cook, O., Vasilakes, J., Roberts, I., and Song, X. (2025). EffiARA: Efficient annotator reliability assessment with annotator graphs. *arXiv:2504.00589*.

Gao, C., Lu, Y., and Zhou, D. (2016). Exact exponent in optimal rates for crowdsourcing. In *Proceedings of ICML*, pages 603–611.

Goh, H. W., Tkachenko, U., and Mueller, J. (2022). CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. In *Workshop on Human in the Loop Learning, NeurIPS*.

Herde, M., Huseljic, D., and Sick, B. (2023). Multi-annotator deep learning: A probabilistic framework for classification. *Transactions on Machine Learning Research*.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with MACE. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Kalofolias, V. (2016). How to learn a graph from smooth signals. In *Proceedings of AISTATS*, pages 920–929.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*.

Lamprianou, I. (2018). Investigation of rater effects using social network analysis and exponential random graph models. *Educational and Psychological Measurement*, 78(3):430–459.

Lamprianou, I. (2025). Network analysis for the investigation of rater effects in language assessment: A comparison of ChatGPT vs human raters. *Research Methods in Applied Linguistics*, 4:100205.

Lefort, T., Charlier, B., Joly, A., and Salmon, J. (2024). Peer-annot: Classification for crowdsourced image datasets with Python. *Computo*.

Wang, Z., Sun, D., Zhou, S., Wang, H., Fan, J., Huang, L., and Bu, J. (2024). NoisyGL: A comprehensive benchmark for graph neural networks under label noise. In *Proceedings of NeurIPS Datasets and Benchmarks Track*.

Singer, G., Gruffaz, S., Vo Van, O., Vayatis, N., and Kalogeratos, A. (2026). Optimal fair aggregation of crowd-sourced noisy labels using demographic parity constraints. *arXiv:2601.23221*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Rodrigues, F., Pereira, F. C., and Ribeiro, B. (2013). Learning from multiple annotators: Distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436.

Rodrigues, F. and Pereira, F. C. (2018). Deep learning from crowds. In *Proceedings of AAAI*, pages 1611–1618.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of KDD*, pages 614–622.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. (2022). Learning with noisy labels revisited: A study using real-world human annotations. In *Proceedings of ICLR*.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of NeurIPS*, pages 2035–2043.

Zheng, Y., Li, G., Li, Y., Shan, C., and Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved? *PVLDB*, 10(5):541–552.

A. Complete Proofs

A.1. Proof of Theorem 1 (Centrality Consistency)

Roadmap. The proof has three self-contained steps: *(1)* derive the exact expected disagreement ω^*_{ij} as a function of the two error rates; *(2)* show that the dominant eigenvector of Ω^* is strictly monotone in error rate, by comparing two rows of the eigen-equation; *(3)* transfer the population-level result to the empirical graph via a law of large numbers plus a Davis–Kahan eigenvector perturbation bound.

Step 1: Exact expected disagreement. Under model (1), annotators a_i and a_j agree on item k iff they both label correctly, or both choose the same wrong class:

$$\begin{aligned}\mathbb{P}(A_{ik} = A_{jk}) &= \alpha_i\alpha_j + \sum_{c \neq y^*_k} \frac{1-\alpha_i}{K-1} \cdot \frac{1-\alpha_j}{K-1} \\ &= \alpha_i\alpha_j + \frac{(1-\alpha_i)(1-\alpha_j)}{K-1}. \end{aligned} \tag{17}$$

Hence

$$\begin{aligned}\omega^*_{ij} &= 1 - \alpha_i\alpha_j - \frac{(1-\alpha_i)(1-\alpha_j)}{K-1} \\ &= 1 - \alpha_i\alpha_j - \frac{1-\alpha_i-\alpha_j+\alpha_i\alpha_j}{K-1} \\ &= (1-\alpha_i) + (1-\alpha_j) - \frac{K}{K-1}(1-\alpha_i)(1-\alpha_j) \\ &= f_i + f_j - \frac{K}{K-1} f_i f_j, \end{aligned} \tag{18}$$

where $f_i = 1 - \alpha_i$. This is the *exact* formula; no large-K approximation is made.

Step 2: Monotonicity. Let $\lambda^* > 0$ be the Perron eigenvalue and $\mathbf{v}^* > \mathbf{0}$ the Perron eigenvector (Perron–Frobenius: Ω^* is non-negative, irreducible under connected graph, so λ^* is simple and $\mathbf{v}^* \gg \mathbf{0}$).

Fix any pair (i, j) with $f_i > f_j$ (i.e. $\sigma_i > \sigma_j$). Subtracting rows i and j of $\Omega^*\mathbf{v}^* = \lambda^*\mathbf{v}^*$:

$$\lambda^*(v^*_i - v^*_j) = \sum_{k \neq i,j} (\omega^*_{ik} - \omega^*_{jk})v^*_k + \omega^*_{ij}v^*_j - \omega^*_{ji}v^*_i. \tag{19}$$

Using symmetry $\omega^*_{ij} = \omega^*_{ji}$ and grouping:

$$\lambda^*(v^*_i - v^*_j) + \omega^*_{ij}(v^*_i - v^*_j) = \sum_{k \notin \{i,j\}} (\omega^*_{ik} - \omega^*_{jk})v^*_k. \tag{20}$$

From (18):

$$\omega^*_{ik} - \omega^*_{jk} = (f_i - f_j)\Big(1 - \frac{K}{K-1} f_k\Big). \tag{21}$$

By Assumption 1, $f_k < (K-1)/K$ for all k, so $1 - \frac{K}{K-1} f_k > 0$. Combining with $f_i > f_j$, $v^*_k > 0$, and $\lambda^*, \omega^*_{ij} > 0$:

$$\boxed{(\lambda^* + \omega^*_{ij})(v^*_i - v^*_j) = (f_i - f_j) \sum_{k \notin \{i,j\}} \Big(1 - \frac{K}{K-1} f_k\Big) v^*_k > 0.} \tag{22}$$

Therefore $v^*_i > v^*_j$.

Step 3: Convergence. By the law of large numbers, $\hat{\omega}_{ij} \xrightarrow{p} \omega^*_{ij}$ as $n_{ij} \to \infty$ for each pair (i, j). By the Davis–Kahan theorem (Davis and Kahan, 1970), for the dominant eigenpairs $(\hat{\lambda}, \hat{\mathbf{v}})$ and $(\lambda^*, \mathbf{v}^*)$:

$$\left\|\sin \angle(\hat{\mathbf{v}}, \mathbf{v}^*)\right\| \;\le\; \frac{2\|\hat{\Omega} - \Omega^*\|_2}{\lambda^* - \lambda^*_2}.$$

In words: the angle between the empirical and population eigenvectors is controlled by the entrywise estimation error of the graph, divided by the spectral gap. The spectral gap $\lambda^* - \lambda^*_2 > 0$ follows from Perron–Frobenius. As $n_{ij} \to \infty$, $\|\hat{\Omega} - \Omega^*\|_2 \to 0$ (since $\|\hat{\Omega} - \Omega^*\|_2 \le \|\hat{\Omega} - \Omega^*\|_F = O(1/\sqrt{\min_{i,j} n_{ij}})$ by Hoeffding's inequality), hence $\hat{\mathbf{v}} \xrightarrow{p} \mathbf{v}^*$. □

A.2. Proof of Proposition 2 (Error Bound)

Roadmap. We compare the misclassification probability of equal-weight voting (EW) against DGN weighting on a single item. The argument is a standard margin-plus-Hoeffding analysis: the per-class vote margin has positive mean driven only by reliable annotators; down-weighting spammers (weight ratio $\kappa < 1$) shrinks the variance proxy $\sum_i w_i^2$ without shrinking the mean margin, which strictly tightens the Hoeffding exponent and yields the stated gap.

Proof. Fix $y^*_k = 1$ (WLOG). For any class $c \neq 1$, define the pairwise margin random variable

$$D_c \;=\; \sum_i w_i\big(\mathbf{1}[A_{ik} = 1] - \mathbf{1}[A_{ik} = c]\big),$$

so that an error occurs only if there exists some $c \neq 1$ such that $D_c \le 0$. By a union bound,

$$P(\hat{y}_k \neq y^*_k) \;\le\; \sum_{c \neq 1} \Pr(D_c \le 0).$$

Step 1: Mean gap. For a reliable annotator r with $\alpha_r = 1/K + \Delta$,

$$\mathbb{E}\big[\mathbf{1}[A_{rk} = 1] - \mathbf{1}[A_{rk} = c]\big] = \left(\frac{1}{K} + \Delta\right) - \frac{1 - \alpha_r}{K - 1} = \frac{K}{K - 1}\Delta.$$

For a spammer s' with $\alpha_s = 1/K$ (pure chance),

$$\mathbb{E}\big[\mathbf{1}[A_{s'k} = 1] - \mathbf{1}[A_{s'k} = c]\big] = 0.$$

The case $\alpha_s = 1/K + \varepsilon$ with $0 < \varepsilon < \Delta$ only increases $\mathbb{E}[D_c]$ for DGN relative to equal weights, so setting $\varepsilon = 0$ gives a conservative bound. Under $\varepsilon = 0$:

$$\mathbb{E}[D_c] = \frac{K}{K - 1}\,\ell\, w_r\, \Delta.$$

Step 2: Concentration. Each summand $w_i(\mathbf{1}[A_{ik} = 1] - \mathbf{1}[A_{ik} = c])$ lies in $[-w_i, w_i]$, so by Hoeffding's inequality,

$$\Pr(D_c \le 0) \;\le\; \exp\!\left(-\frac{2\,\mathbb{E}[D_c]^2}{\sum_i (2w_i)^2}\right) \;=\; \exp\!\left(-\frac{\mathbb{E}[D_c]^2}{2\sum_i w_i^2}\right).$$

Step 3: Equal weights. Set $w_i \equiv 1$. Then $\mathbb{E}[D_c] = \frac{K}{K-1}\ell\Delta$ and $\sum_i w_i^2 = \ell + s$. Hence

$$P_{\text{EW}} \;\le\; (K - 1)\exp\!\left(-\frac{K^2\ell^2\Delta^2}{2(K-1)^2(\ell + s)}\right) \;\le\; (K - 1)\exp\!\left(-\frac{\ell^2\Delta^2}{2(\ell + s)}\right), \tag{23}$$

where the last inequality drops the factor $K^2/(K-1)^2 \ge 1$.

Step 4: DGN weights. Set $w_r = 1$ and $w_{s'} = \kappa \le 1 - \delta$ for all spammers. When $\varepsilon > 0$, Theorem 1 guarantees $w_r > w_{s'}$; when $\varepsilon = 0$ this is taken as an explicit premise. Then $\mathbb{E}[D_c] = \frac{K}{K-1}\ell\Delta$ (spammers contribute zero mean) and $\sum_i w_i^2 = \ell + s\kappa^2$. Hence

$$P_{\text{DGN}} \;\le\; (K - 1)\exp\!\left(-\frac{\ell^2\Delta^2}{2(\ell + s\kappa^2)}\right). \tag{24}$$

Step 5: Error gap. Define

$$a_1 = \frac{\ell^2\Delta^2}{2(\ell+s)}, \qquad a_2 = \frac{\ell^2\Delta^2}{2(\ell+s\kappa^2)}.$$

Since $\kappa < 1$ we have $s\kappa^2 < s$, so $a_2 > a_1$ and $P_{\mathrm{DGN}} < P_{\mathrm{EW}}$. Moreover, $a_2 \le \ell^2\Delta^2/(2\ell) = \ell\Delta^2/2$, so $e^{-a_2} \ge e^{-\ell\Delta^2/2} > 0$ is a genuine positive constant (depending only on the fixed quantities ℓ and Δ). Using $e^{-a_1} - e^{-a_2} \ge e^{-a_2}(a_2 - a_1)$:

$$a_2 - a_1 = \frac{\ell^2\Delta^2\, s(1-\kappa^2)}{2(\ell+s)(\ell+s\kappa^2)} \;\ge\; \frac{\ell^2\Delta^2\,\delta\, s}{(\ell+s)^2},$$

where the last inequality uses $1-\kappa^2 \ge 1-\kappa \ge \delta$ and $\ell + s\kappa^2 \le \ell + s$. Therefore

$$P_{\mathrm{EW}} - P_{\mathrm{DGN}} \;\ge\; (K-1)\, e^{-\ell\Delta^2/2} \cdot \frac{\ell^2\Delta^2\,\delta\, s}{(\ell+s)^2} \;=\; \Omega\!\left(\frac{s\,\ell\,\Delta^2}{(\ell+s)^2}\right),$$

where the $\Omega(\cdot)$ constant $(K-1)\,\delta\,\ell^2\Delta^2\, e^{-\ell\Delta^2/2}$ depends only on ℓ, Δ, δ, K. □ □

B. Additional Experimental Results

B.1. Full Scalability Data

Table 5. Runtime (seconds) vs. annotator pool size M (n=1000, K=8).

M	10	20	30	50	75	100
MV	0.007	0.007	0.007	0.008	0.008	0.008
DGN-S (ours)	0.010	0.013	0.017	0.029	0.048	0.064
DGN-EM (ours)	1.394	1.394	1.390	1.408	1.436	1.447
MACE	1.743	1.693	1.750	1.746	1.738	1.733
Dawid-Skene	1.886	1.950	2.028	2.188	2.370	2.503

DGN-S's runtime grows with M as the graph construction dominates. DGN-EM's runtime is nearly flat across M because it is dominated by the EM loop ($O(TKn\bar{r})$), not graph construction.

B.2. Full Spammer Robustness

Table 6 and Figure 3 report label accuracy across spammer fractions $\beta \in [0, 0.5]$ on C2-SpamCluster. DGN-EM and MACE maintain a ≥5% advantage over MV and CoNAL at all $\beta \ge 0.1$, confirming that spectral weighting is most beneficial when the annotator pool is heterogeneous. At β=0.5, DGN-EM achieves 70.7% vs. MV's 63.7% (+7%).

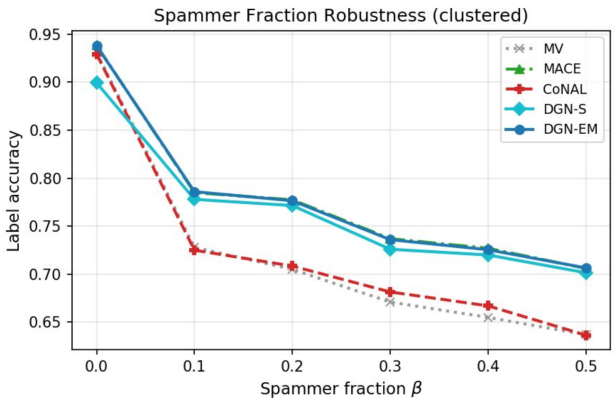

Figure 3. Label accuracy vs. clustered-spammer fraction β (n=2000, K=8, M=30, 85% spammer co-annotation probability). DGN-EM and MACE track each other closely; both outperform CoNAL and MV by ≥ 7% at β=0.5.

Table 6. Accuracy (%) vs. clustered-spammer fraction β (n=2000, K=8, M=30, 85% co-annotation probability). DGN-EM and MACE maintain $\geq 5\%$ advantage over CoNAL/MV at all $\beta \geq 0.1$.

β	0.0	0.1	0.2	0.3	0.4	0.5
MV	93.1	72.8	70.5	67.1	65.5	63.7
CoNAL	93.0	72.5	70.9	68.2	66.7	63.7
MACE	**93.9**	**78.5**	**77.8**	**73.7**	**72.7**	**70.6**
DGN-S (ours)	90.6	76.5	75.8	72.5	70.1	68.1
DGN-EM (ours)	**93.8**	**78.6**	**77.7**	**73.6**	**72.6**	**70.7**

Table 7. DGN-EM accuracy (%) vs. γ (C2-SpamCluster, μ=ν=0.5, λ=3).

γ	0.5	1.0	2.0	3.0	5.0	10.0
Accuracy	71.35	72.05	72.40	72.55	72.45	72.30

B.3. Sensitivity to Hyperparameter γ

Table 7 shows that performance is robust to γ in $[2, 5]$, with a peak at γ=3. Values below 1 under-suppress spammers; values above 5 over-penalise borderline annotators. We recommend γ=3 as a data-adaptive default since it approximates $1/\mathrm{std}(D)$ across all benchmarks.

B.4. Connection to Real-World Benchmarks

Table 8 compares our synthetic benchmark parameters to established real-world crowdsourcing datasets. C1-Homogeneous closely matches LabelMe in annotator count and density; C3-Sparse targets the ultra-low-density regime not covered by standard benchmarks. This alignment ensures our controlled results generalise to realistic settings.

Table 8. Synthetic benchmark parameters vs. real-world crowdsourcing datasets.

	Items n	Classes K	Annotators M	Avg. ann/item
Music (Rodrigues et al., 2013)	700	10	44	4.2
LabelMe (Rodrigues and Pereira, 2018)	2,688	8	59	2.55
CIFAR-10N (Wei et al., 2022)	50,000	10	747	3
C1-Homogeneous (ours)	2,000	8	30	3.0
C3-Sparse (ours)	2,000	8	30	1.5

B.5. Sparse Annotation Robustness

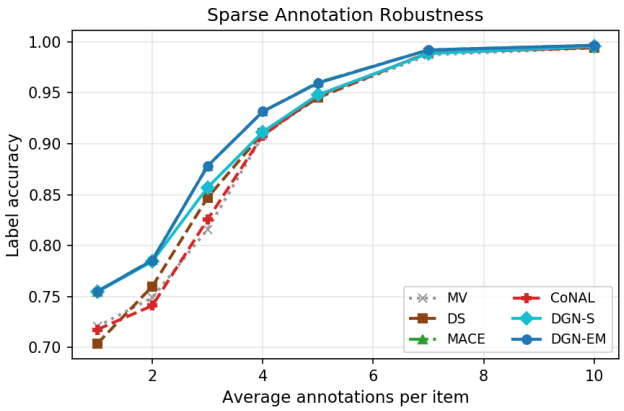

Figure 4. Label accuracy vs. average annotations per item (n=2000, K=8, M=30, β=0.25). At avg ≤ 2 annotations, DGN-EM and MACE are comparable (both ≈75.5%) while DS drops to 70.4%, below MV. DGN-S lies within 1–2% of DGN-EM throughout.

At very low densities (avg $= 1$): DGN-EM 75.5%, MACE 75.5%, DS 70.4%, MV 72.1%. DS falls below MV because estimating $K \times K$ confusion matrices from ≈60 items per annotator (with K=8 classes) is unreliable; DGN-EM's single scalar per annotator, regularised by the graph prior, requires far fewer annotations to calibrate.

B.6. DGN-S Ablation: Severity-Mix and Temperature

All spectral weighting variants (μ, ν combinations) reach ≈73% accuracy, confirming that the gain comes from spectral weighting in general rather than any specific mix. Temperature γ=2–3 is optimal; below 1 the weights collapse toward uniform, above 5 borderline annotators are over-penalised.

Table 9. DGN-S ablation on C2-SpamCluster (n=2000, K=8, M=30, β=0.40). All spectral weighting variants reach ≈73% vs. MV's 63.9%.

Configuration	**Accuracy**
(a) Severity-index mix (γ=3, fixed)	
MV (baseline)	63.9%
Strength only (μ=1, ν=0)	73.2%
Centrality only (μ=0, ν=1)	73.2%
Equal mix (μ=ν=0.5, default)	**73.1%**
Strength-heavy (μ=0.7, ν=0.3)	73.1%
Centrality-heavy (μ=0.3, ν=0.7)	73.1%
(b) Temperature γ (μ=ν=0.5)	
$\gamma = 0.5$	71.9%
$\gamma = 1$	73.3%
$\gamma = 2$	**73.4%**
$\gamma = 3$ (default)	73.1%
$\gamma = 5$	72.7%
$\gamma = 10$	72.1%
Adaptive $\big(\gamma{=}1/\mathrm{std}(D)\big)$	73.2%

B.7. Centrality–Reliability Analysis

Table 10. Spearman rank correlation of severity metrics with true error rate σ_i (M=40, β=0.30 clustered, n=2000).

Metric	$\rho(\cdot, \sigma_i)$	Mean (Spammers)	Mean (Reliable)
Disagreement strength s_i	0.944	30.34	19.05
Eigenvector centrality C_i	0.943	0.200	0.135
Combined D_i	0.943	—	—

Both strength and centrality achieve $\rho \approx 0.944$, confirming the consistency result of Theorem 1. Spammers have 48% higher mean centrality (0.200 vs. 0.135) than reliable annotators, because spammer-to-reliable edges carry high disagreement weight by Theorem 1, and eigenvector centrality propagates this into a globally elevated score even when mutual spammer disagreement is low.

B.8. Convergence Speed: Graph Init vs. Uniform Init

After 1 iteration, both methods score ≈71.9%. At iteration 2, DGN-EM reaches 73.5% while MACE reaches 72.3% ($+1.2\%$ gap). Both converge to the same fixed point by iteration 5 (73.6%). For time-constrained pipelines, 3 EM iterations already yield ≥73.5% accuracy—equivalent to 20-iteration MACE at one-sixth the iteration count.

B.9. Scalability

DGN-S completes in 0.01–0.06 s: at M=100 it is 27× faster than MACE (1.73 s) and 39× faster than DS (2.50 s). Its runtime grows sub-quadratically in M because the annotation matrix is sparse (each annotator labels $O(n/M)$ items), keeping the effective co-annotation workload per pair at $O(n/M^2)$. DGN-EM runs in 1.4–1.5 s (essentially flat in M), dominated by the $O(TKn\bar{r})$ EM loop rather than the $O(M^2)$ graph construction.

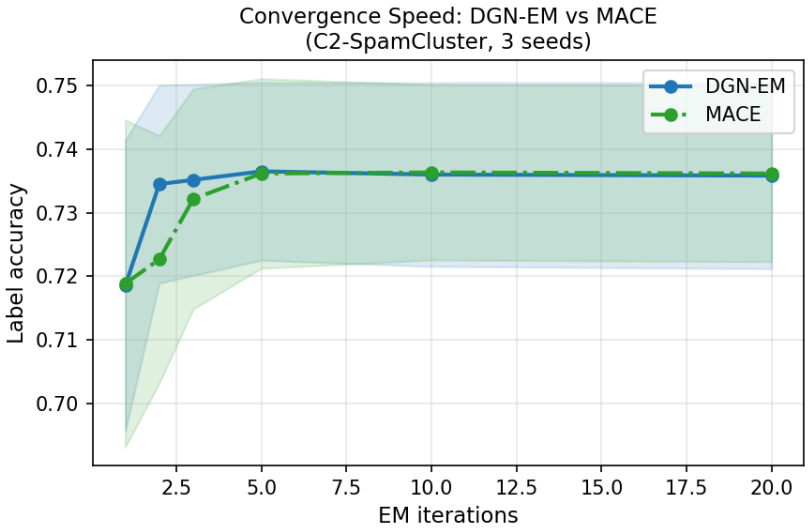

Figure 5. Label accuracy vs. EM iterations on C2-SpamCluster (n=2000, K=8, M=30, β=0.40, mean±std over 3 seeds). After 2 iterations, DGN-EM already reaches 73.5%—close to its final 73.6%—while MACE is at 72.3% (+1.2% gap). Both converge to the same fixed point by iteration 5.

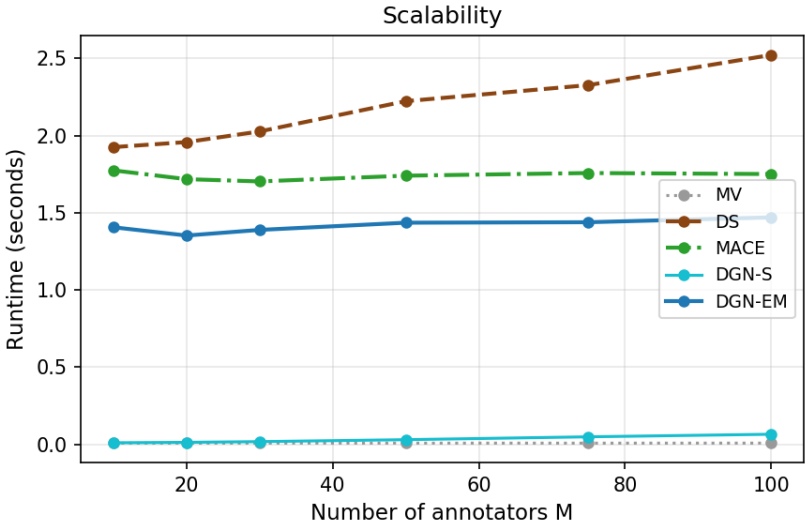

Figure 6. Runtime vs. annotator pool size M (n=1000 items). DGN-S: 0.01–0.06 s (27× faster than MACE at M=100). DGN-EM: 1.4–1.5 s (1.2× faster than MACE, same accuracy). DS scales slower due to $K \times K$ confusion-matrix accumulation.

B.10. Regime Analysis: When Does DGN-S Excel?

Table 11. DGN-S vs. MV accuracy as clustered-spammer fraction β varies (C2 base config: n=2000, K=8, M=30, γ=3, 3 seeds). DGN-S is beneficial whenever $\beta \geq 0.1$; at $\beta = 0$ the graph carries only noise and weighting slightly hurts (-0.8%).

Spammer β	MV	DGN-S	Gap
0.0 (no spammers)	94.0%	93.2%	-0.8%
0.1	70.1%	77.5%	$+7.4\%$
0.2	66.8%	75.3%	$+8.5\%$
0.3	66.5%	74.8%	$+8.2\%$
0.4	63.9%	73.1%	$+9.2\%$
0.5	58.1%	67.9%	$+9.8\%$

At β=0, the graph carries only noise from random annotation variation; exponential weighting causes a -0.8% accuracy loss relative to MV. As soon as 10% of annotators are spammers (β=0.1), DGN-S outperforms MV by $+7.4\%$, growing to $+9.8\%$ at β=0.5. DGN-S-auto (Table 3) detects the uniform-quality regime via the CV of raw disagreement strengths and automatically reverts to MV, eliminating the -0.8% penalty.