Uncertainty Aware Semi-Supervised Learning on Graph Data

Xujiang Zhao[1], **Feng Chen**[1], **Shu Hu**[2], **Jin-Hee Cho**[3]
[1]The University of Texas at Dallas, {xujiang.zhao, feng,chen}@utdallas.edu
[2]University at Buffalo, SUNY, shuhu@buffalo.edu
[3]Virginia Tech, jicho@vt.edu

Abstract

Thanks to graph neural networks (GNNs), semi-supervised node classification has shown the state-of-the-art performance in graph data. However, GNNs have not considered different types of uncertainties associated with class probabilities to minimize risk of increasing misclassification under uncertainty in real life. In this work, we propose a multi-source uncertainty framework of GNN that reflecting various types of uncertainties in both deep learning and belief/evidence theory domains for node classification predictions. By collecting evidence from the given labels of training nodes, the *Graph-based Kernel Dirichlet distribution Estimation* (GKDE) method is designed for accurately predicting Dirichlet distributions and detecting out-of-distribution (OOD). We validated the outperformance of the proposed model compared to the state-of-the-art counterparts in terms of misclassification detection and OOD detection based on six real network datasets. We found that dissonance-based detection yielded the best results on misclassification detection while vacuity-based detection was best for OOD detection. To clarify the reasons behind the results, we provided the theoretical proof that explains the relationships between different types of uncertainty considered in this work.

1 Introduction

Inherent uncertainties derived with different root causes have emerged as serious hurdles to find effective solutions for real world problems. Critical safety concerns have been brought due to lack of considering diverse causes of uncertainties, resulting in high risk due to misinterpretation of uncertainties (e.g., misdetection or misclassification of an object by an autonomous vehicle). Graph neural networks (GNNs) [12, 22] have received tremendous attention in the data science community. Despite their superior performance in semi-supervised node classification/regression, they didn't consider various types of uncertainties in the decision process of node classification/regression. Predictive uncertainty estimation [11] using Bayesian NNs (BNNs) has been explored for classification prediction/regression in the computer vision applications, with well-known uncertainties, aleatoric uncertainty (AU) and epistemic uncertainty (EU). AU refers to data uncertainty from statistical randomness (e.g., inherent noises in observations) while EU indicates model uncertainty due to limited knowledge/ignornace in collected data. In the belief/evidence theory domain, Subjective Logic (SL) [9] considered vacuity (or a lack of evidence) as uncertainty in a subjective opinion. Recently other uncertainty types, such as dissonance, consonance, vagueness, and monosonance [9], have been discussed based on SL to measure them based on their different root causes.

We first considered multidimensional uncertainty types in both deep learning (DL) and belief theory domains to predict node classification and out-of-distribution (OOD) detection. We performed missclassification and OOD detection based on semi-supervised node classification. By leveraging the learning capability of GNNs and considering multidimensional uncertainties, we propose a

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

uncertainty framework that can simultaneously estimate different uncertainty types associated with the predicted class probabilities. This work has the following **key contributions**:

- **A multi-source uncertainty framework for GNNs**. Our proposed framework first provides the estimation of various types of uncertainty from both DL and evidence/belief theory domains, such as dissonance (derived from conflicting evidence) and vacuity (derived from lack of evidence). In addition, we designed a Graph-based Kernel Dirichlet distribution Estimation (GKDE) method to reduce errors in uncertainty prediction.
- **Theoretical analysis**: Our work is the first that provides a theoretical analysis about the relationships between different types of uncertainties considered in this work. We provide a theoretical guarantee that an OOD node may have a high uncertainty under GKDE.
- **Comprehensive experiments for validating the performance of our proposed framework**: Based on the six real graph datasets, we compared the performance of our proposed framework with that of other competitive counterparts. We found that the dissonance-based detection yielded the best results on misclassification detection while vacuity-based detection was best for OOD detection.

2 Related Work

DL research has mainly considered *aleatoric* uncertainty (AU) and *epistemic* uncertainty (EU) using BNNs for computer vision applications. AU consists of homoscedastic uncertainty (i.e., constant errors for different inputs) and heteroscedastic uncertainty (i.e., different errors for different inputs) [4]. A Bayesian DL framework was presented to simultaneously estimate both AU and EU in regression (e.g., depth regression) and classification (e.g., semantic segmentation) tasks [11]. Later, *distributional uncertainty* was defined based on distributional mismatch between testing and training data distributions [14]. *Dropout variational inference* [5] was used for an approximate inference in BNNs using epistemic uncertainty. Other algorithms have considered overall uncertainty in node classification [3, 13, 23]; but no prior work has considered uncertainty decomposition in GNNs.

In the belief/evidence theory domain, uncertainty reasoning has been substantially explored, such as Fuzzy Logic [1], Dempster-Shafer Theory (DST) [19], or Subjective Logic (SL) [8]. Belief theory focuses on reasoning inherent uncertainty in information resulting from unreliable, incomplete, deceptive, and/or conflicting evidence. SL considered uncertainty in subjective opinions in terms of *vacuity* (i.e., a lack of evidence) and *vagueness* (i.e., failing in discriminating a belief state) [8]. Recently, other uncertainty types have been studied, such as *dissonance* caused by conflicting evidence[9]. In deep NNs, [18] proposed EDL model, using SL to train a deterministic NN for supervised classification in computer vision based on sum of squared loss. However, EDL didn't consider a general method of estimating multidimensional uncertainty or graph structure.

3 Multidimensional Uncertainty and Subjective Logic

This section provides the overview of SL and discusses SL-based multiple types of uncertainties, called *evidential uncertainty*, with the measures of *vacuity* and *dissonance*. In addition, we give a brief overview of *probabilistic uncertainty*, discussing the measures of *aleatoric* and *epistemic*.

3.1 Subjective Logic

SL offers the formulation of a subjective opinion based on both probabilistic logic (PL) [15] and belief theory (BT) [20] with two unique extensions. First, SL explicitly represents uncertainty by introducing vacuity of evidence (or uncertainty mass) in its opinion representation. This addresses the limitations of PL by modeling a lack of confidence in probabilities. Second, SL extends the traditional BT by incorporating base rates as the prior probabilities in Bayesian theory. The Bayesian nature of SL allows it to use second-order uncertainty to express and reason the uncertainty mass, where second-order uncertainty is represented in terms of a probability density function (PDF) over first-order probabilities [8]. For multi-class problems, we use a multinomial distribution (first-order uncertainty) to model class probabilities and a Dirichlet PDF (second-order uncertainty) to model the distribution of class probabilities. Second-order uncertainty enriches uncertainty representation with evidence information, playing a key role in distinguish OOD from conflict prediction as detailed later.

Opinions are the arguments in SL. In the multi-class setting, the multinomial opinion of a random variable y in domain $\mathbb{Y} = \{1, ..., K\}$ is given by a triplet as:

$$\omega = (\boldsymbol{b}, u, \boldsymbol{a}), \text{ with } \sum_{k=1}^{K} b_k + u = 1 \tag{1}$$

where $\boldsymbol{b} = (b_1, ..., b_K)^T, u$, and $\boldsymbol{a} = (a_1, ..., a_K)^T$ denote the belief mass distribution over $\mathbb{Y}$, uncertainty mass representing vacuity of evidence, and base rate distribution over $\mathbb{Y}$, respectively, and $\forall k, a_k \geq 0, b_k \geq 0, u \geq 0$. The probability that y is assigned to the k-class is given by $P(y = k) = b_k + a_k u$, which combines the belief mass with the uncertain mass using the base rates. In the multi-class setting, a_k can be regarded as the prior preference over the k-th class. When no specific preference is given, we assign all the base rates as $1/K$.

3.2 Evidential Uncertainty

In this section, we explain how the second order uncertainty (evidential uncertainty) is derived from the first order uncertainty as a Dirichlet PDF. Given a set of random variables $\mathbf{p} = (p_1, ..., p_K)^T$, where $\mathbf{p}$ is distributed on a simplex of dimensionality $K - 1$, a conditional distribution $P(y = k|\mathbf{p}) = p_k$ can be represented by the marginal distribution, $P(y) = \int P(y|\mathbf{p})p(\mathbf{p})d\mathbf{p}$. We define $p(\mathbf{p})$ as a Dirichlet PDF over $\mathbf{p}$: $\text{Dir}(\mathbf{p}|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T$ is a K-dimensional strength vector, with $\alpha_k \geq 0$ denoting the effective number of observations of the k-th class. SL explicitly introduces uncertain evidence through a weight W representing non-informative evidence and redefines the strength parameter as: $\alpha_k = e_k + a_k W$, with $e_k \geq 0, \forall k \in \mathbb{Y}$, where e_k is the amount of evidence (or the number of observations) to support the k-th class and W is usually set to K, i.e., the number of classes. Given the new definition of the strength parameter, the expectation of the class probabilities $\mathbf{p} = (p_1, ..., p_K)^T$ is given by:

$$\mathbb{E}[p_k] = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j} = \frac{e_k + a_k W}{\sum_{j=1}^{K} e_j + W} \tag{2}$$

where $a_k = 1/K$. By marginalizing out $\mathbf{p}$, we can derive an evidence-based expression of belief mass and uncertainty mass:

$$b_k = \frac{e_k}{S} \quad \forall k \in \mathbb{Y}, \quad u_v = \frac{W}{S}, \text{ with } S = \sum_{k=1}^{K} \alpha_k \tag{3}$$

SL categorizes uncertainty into two primary sources [8]: (1) basic belief uncertainty that results from specific aspects of belief mass in isolation and (2) intra-belief uncertainty that results from the relationships between belief masses and uncertainty mass. As a result, these two sources of uncertainty boil down to ***vacuity*** and ***dissonance***, respectively, that correspond to vacuous belief and contradicting beliefs. In particular, vacuity of an opinion ω is captured by uncertainty mass u_v, which is defined in (3) and dissonance of an opinion [9] is defined as:

$$diss(\omega) = \sum_{k=1}^{K} \Big(\frac{b_k \sum_{j \neq k} b_j \text{Bal}(b_j, b_k)}{\sum_{j \neq k} b_j} \Big), \text{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k} & \text{if } b_i b_j \neq 0 \\ 0 & \text{if } \min(b_i, b_j) = 0 \end{cases} \tag{4}$$

where $\text{Bal}(b_j, b_k)$ is the relative mass balance function between two belief masses. The belief dissonance of an opinion is measured based on how much belief supports individual classes. Consider a binary classification example with a binomial opinion given by $(b_1, b_2, u, \boldsymbol{a}) = (0.49, 0.49, 0.02, \boldsymbol{a})$. Based on (4), it has a dissonance value of 0.98. In this case, although the vacuity is close to zero, a high dissonance indicates that one cannot make a clear decision because both two classes have the same amount of supporting both beliefs and hence strongly conflict with each other.

3.3 Probabilistic Uncertainty

For classification, the estimation of the probabilistic uncertainty relies on the design of an appropriate Bayesian DL model with parameters $\boldsymbol{\theta}$. Given input x and dataset $\mathcal{G}$, we estimate a class probability by $P(y|x) = \int P(y|x; \boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{G})d\boldsymbol{\theta}$, and obtain ***epistemic uncertainty*** estimated by mutual information [2, 14]:

$$\underbrace{I(y, \boldsymbol{\theta}|x, \mathcal{G})}_{\textbf{\textit{Epistemic}}} = \underbrace{\mathcal{H}\big[\mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{G})}[(y|x; \boldsymbol{\theta})]\big]}_{\textbf{\textit{Entropy}}} - \underbrace{\mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{G})}\big[\mathcal{H}(y|x; \boldsymbol{\theta})\big]}_{\textbf{\textit{Aleatoric}}}, \tag{5}$$

where $\mathcal{H}(\cdot)$ is Shannon's entropy of a probability distribution. The first term indicates ***entropy***, representing the total uncertainty, and the second term is ***aleatoric***, indicating data uncertainty. By computing the difference between entropy and aleatoric uncertainties, we obtain epistemic uncertainty, which also refers to uncertainty from model parameters.

4 Relationships Between Multiple Uncertainties

We use the shorthand notations u_v, u_{diss}, u_{alea}, u_{epis}, and u_{en} to represent vacuity, dissonance, aleatoric, epistemic and entropy, respectively.

To interpret multiple types of uncertainty, we show an ideal classification example in Figure 1. For a prediction to confident, the subjective multinomial opinion, following a Dirichlet distribution, will yield a sharp distribution on one corner of the simplex (see Figure 1 (a)). For a conflict prediction (CP) (due to the conflicting evidence), the multinomial opinion should yield a central distribution, representing confidence to predict a flat categorical distribution over class labels (see Figure 1 (b)). For an OOD (no evidence, i.e, $\alpha = [1, 1, 1]$) scenario, a multinomial opinion would yield a flat distribution over the simplex (Figure 1c), indicating high uncertainty due to the absence of evidence. The first technical contribution of this work is as follows.

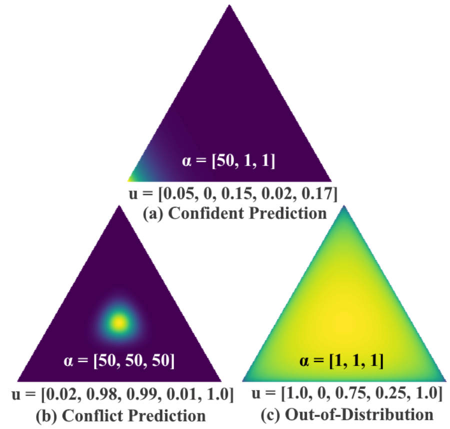

Figure 1: Multiple uncertainties of different prediction. Let $\mathbf{u} = [u_v, u_{diss}, u_{alea}, u_{epis}, u_{en}]$.

Theorem 1 *Given a total Dirichlet strength $S = \sum_{i=1}^{K} \alpha_i$, where K is the number of classes, for any opinion ω on a multinomial random variable y, we have*

1. *General relations on all prediction scenarios.*
   (a) $u_v + u_{diss} \le 1$;
   (b) $u_v > u_{epis}$
2. *Special relations on the OOD and the CP.*
   (a) *For an out-of-distribution sample with uniform prediction, i.e., $\alpha = [1, \ldots, 1]$, we have*
   $$1 = u_v = u_{en} > u_{alea} > u_{epis} > u_{diss} = 0$$
   (b) *For an in-distribution sample with conflict prediction, i.e., $\alpha = [\alpha_1, \ldots, \alpha_K]$ with $\alpha_1 = \alpha_2 = \cdots = \alpha_K$, if $S \to \infty$, we have $u_{en} > u_{alea} > u_{diss} > u_v > u_{epis}$ with*
   $$u_{en} = 1, u_{diss} \to 1, u_{alea} \to 1, u_v \to 0, u_{epis} \to 0$$

The proof of Theorem 1 can be found in Appendix A.1. As demonstrated in Theorem 1 and Figure 1, entropy can not distinguish OOD (Figure 1 (c)) and conflict predictions (Figure 1 (b)) because entropy is high for both cases. Similarly, neither aleatoric uncertainty nor epistemic uncertainty can distinguish OOD from conflict predictions as well. That is, in both cases, aleatoric uncertainty is high while the epistemic uncertainty is low. On the other hand, vacuity and dissonance can clearly distinguish OOD from conflict prediction. For example, OOD objects typically show high vacuity with low dissonance while conflict predictions exhibit low vacuity with high dissonance. This observation is confirmed through the empirical validation via our extensive experiments in terms of misclassification and OOD detection tasks.

5 Uncertainty-Aware Semi-Supervised Learning

In this section, we describe our proposed uncertainty framework based on semi-supervised node classification problem (see Figure 2).

5.1 Problem Definition

Given an input graph $\mathcal{G} = (\mathbb{V}, \mathbb{E}, \mathbf{r}, \mathbf{y}_{\mathbb{L}})$, where $\mathbb{V} = \{1, \cdots, N\}$ is a ground set of nodes, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ is a ground set of edges, $\mathbf{r} = [\mathbf{r}_1, \cdots, \mathbf{r}_N]^T \in \mathbb{R}^{N \times d}$ is a node-level feature matrix, $\mathbf{r}_i \in \mathbb{R}^d$ is the feature vector of node i, $\mathbf{y}_{\mathbb{L}} = \{y_i \mid i \in \mathbb{L}\}$ are the labels of the training nodes $\mathbb{L} \subset \mathbb{V}$, and $y_i \in \{1, \ldots, K\}$ is the class label of node i. **We aim to predict**: (1) the **class probabilities** of the testing nodes: $\mathbf{p}_{\mathbb{V} \setminus \mathbb{L}} = \{\mathbf{p}_i \in [0, 1]^K \mid i \in \mathbb{V} \setminus \mathbb{L}\}$; and (2) the **associated multidimensional uncertainty estimates** introduced by different root causes: $\mathbf{u}_{\mathbb{V} \setminus \mathbb{L}} = \{\mathbf{u}_i \in [0, 1]^m \mid i \in \mathbb{V} \setminus \mathbb{L}\}$, where $p_{i,k}$ is the probability that the class label $y_i = k$ and m is the total number of uncertainty types.

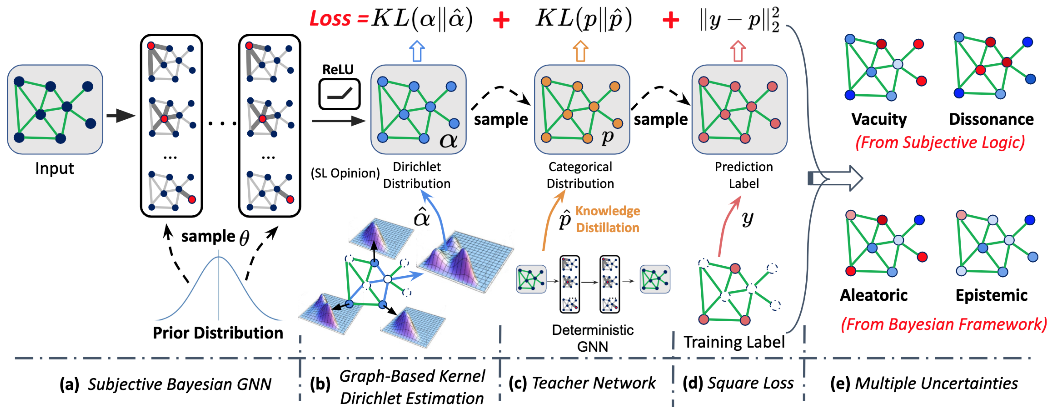

Figure 2: **Uncertainty Framework Overview.** Subjective Bayesian GNN (a) designed for estimating the different types of uncertainties. The loss function includes square error (d) to reduce bias, GKDE (b) to reduce errors in uncertainty estimation and teacher network (c) to refine class probability.

5.2 Proposed Uncertainty Framework

Learning evidential uncertainty. As introduced from Section 3.1, evidential uncertainty can be derived from multinomial opinions or equivalently Dirichlet distributions to model a probability distribution for the class probabilities. Therefore, we design a Subjective GNN (S-GNN) f to form their multinomial opinions for the node-level Dirichlet distribution $\text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i)$ of a given node i. Then, the conditional probability $P(\mathbf{p}|A,\mathbf{r};\boldsymbol{\theta})$ can be obtained by:

$$P(\mathbf{p}|A,\mathbf{r};\boldsymbol{\theta}) = \prod\nolimits_{i=1}^{N} \text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i),\ \boldsymbol{\alpha}_i = f_i(A,\mathbf{r};\boldsymbol{\theta}), \tag{6}$$

where f_i is the output of S-GNN for node i, $\boldsymbol{\theta}$ is the model parameters, and A is an adjacency matrix. The Dirichlet probability function $\text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i)$ is defined by:

$$\text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i) = \frac{\Gamma(S_i)}{\prod_{k=1}^{K}\Gamma(\alpha_{ik})}\prod\nolimits_{k=1}^{K} p_{ik}^{\alpha_{ik}-1}, \tag{7}$$

Note that S-GNN is similar to classical GNN, the only difference is that we use an activation layer (e.g., *ReLU*) instead of the *softmax* layer (only outputs class probabilities) such that S-GNN would output non-negative value, which is taken as the parameters for the predicted Dirichlet distribution.

Learning probabilistic uncertainty. Since probabilistic uncertainty relies on Bayesian framework, we proposed a Subjective Bayesian GNN (S-BGNN) that adapts S-GNN to a Bayesian framework, which considers model parameter θ following a distribution. The joint class probability of $\mathbf{y}$ can be estimated by:

$$\begin{aligned} P(\mathbf{y}|A,\mathbf{r};\mathcal{G}) &= \int\int P(\mathbf{y}|\mathbf{p})P(\mathbf{p}|A,\mathbf{r};\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{G})d\mathbf{p}d\boldsymbol{\theta} \\ &\approx \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\int P(\mathbf{y}_i|\mathbf{p}_i)P(\mathbf{p}_i|A,\mathbf{r};\boldsymbol{\theta}^{(m)})d\mathbf{p}_i, \quad \boldsymbol{\theta}^{(m)} \sim q(\boldsymbol{\theta}) \end{aligned} \tag{8}$$

where $P(\boldsymbol{\theta}|\mathcal{G})$ is the posterior, which can be estimated with dropout inference, providing an approximate solution of posterior $q(\boldsymbol{\theta})$ and taking samples from the posterior distribution of models [5]. Thanks to the benefit of using the dropout inference, **training a DL model directly by minimizing the cross entropy (or square error) loss function can effectively minimize the KL-divergence between the approximated distribution and the full posterior (i.e., KL[$q(\boldsymbol{\theta})\|P(\theta|\mathcal{G})$]) in variational inference** [5, 10]. For interested readers, please refer to more detail in Appendix B.8.

Therefore, training S-GNN with stochastic gradient descent enables learning of an approximated distribution of weights, which can provide good explainability of data and prevent overfitting. We use a *loss function* to compute its Bayes risk with respect to the sum of squares loss $\|\mathbf{y}-\mathbf{p}\|_2^2$ by:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum\nolimits_{i\in\mathbb{L}}\int\|\mathbf{y}_i-\mathbf{p}_i\|_2^2\cdot P(\mathbf{p}_i|A,\mathbf{r};\boldsymbol{\theta})d\mathbf{p}_i = \sum\nolimits_{i\in\mathbb{L}}\sum\nolimits_{k=1}^{K}\big(y_{ik}-\mathbb{E}[p_{ik}]\big)^2 + \text{Var}(p_{ik}), \tag{9}$$

where $\mathbf{y}_i$ is a one-hot vector encoding the ground-truth class with $y_{ij}=1$ and $y_{ik}\neq$ for all $k\neq j$, j is class label. Eq. (9) aims to minimize the prediction error and variance, leading to maximizing the classification accuracy of each training node by removing excessive misleading evidence.

5.3 Graph-based Kernel Dirichlet distribution Estimation (GKDE)

The loss function in Eq. (9) is designed to measure sum of squared loss based on class labels of training nodes. However, it does not directly measure the quality of the predicted node-level Dirichlet distributions. To address this limitation, we proposed *Graph-based Kernel Dirichlet distribution Estimation* (GKDE) to better estimate Dirichlet distribution by using graph structure information. The key idea of GKDE is to estimate prior Dirichlet distribution parameters for each node based on the class labels of training nodes (see Figure 2 (b)). Then, we use the estimated prior Dirichlet distribution in the training process to learn the following patterns: (i) nodes with high vacuity (due to lack of evidence) will be shown far from training nodes; and (ii) nodes with high dissonance (due to conflicting evidence) will be shown in the boundaries of classes.

Based on SL, let each training node represent one evidence for its class label. Denote the contribution of evidence estimation for testing node j from training node i by $\mathbf{h}(y_i, d_{ij}) = [h_1, \ldots, h_k, \ldots, h_K] \in [0, 1]^K$ and $h_k(y_i, d_{ij})$ is obtained by:

$$h_k(y_i, d_{ij}) = \begin{cases} 0 & y_i \neq k \\ g(d_{ij}) & y_i = k \end{cases} \tag{10}$$

where $g(d_{ij}) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{d_{ij}^2}{2\sigma^2})$ is the Gaussian kernel function to estimate the distribution effect between nodes i and j, and d_{ij} means the **node distance** (**a shortest path between nodes** i **and** j), and σ is the bandwidth parameter. The prior evidence estimation based GKDE: $\hat{\boldsymbol{e}}_j = \sum_{i \in \mathbb{L}} \mathbf{h}(y_i, d_{ij})$ where $\mathbb{L}$ is a set of training nodes, and the prior Dirichlet distribution $\hat{\boldsymbol{\alpha}}_j = \hat{\boldsymbol{e}}_j + \mathbf{1}$. During training process, we minimize the KL-divergence between model predictions of Dirichlet distribution and prior distribution: $\min \text{KL}[\text{Dir}(\boldsymbol{\alpha}) \| \text{Dir}(\hat{\boldsymbol{\alpha}})]$. This process can prioritize the extent of data relevance based on estimated evidential uncertainty, which is proven effective based on the proposition below.

Proposition 1 *Given L training nodes, for any testing nodes i and j, let $\boldsymbol{d}_i = [d_{i1}, \ldots, d_{iL}]$ is the graph distances from nodes i to training nodes, and $\boldsymbol{d}_j = [d_{j1}, \ldots, d_{jL}]$ is the graph distances from nodes j to training nodes, where d_{il} is the node distance between nodes i and l. If for all $l \in \{1, \ldots, L\}$, $d_{il} \geq d_{jl}$, then we have*

$$\hat{u}_{v_i} \geq \hat{u}_{v_j}$$

where $\hat{u}_{v_i}$ and $\hat{u}_{v_j}$ are vacuity uncertainties estimated of nodes i and j based on GKDE.

The proof for this proposition can be found in Appendix A.2. The above proposition shows that if a testing node is too far from training nodes, the vacuity increases, implying that an OOD node is expected to have high vacuity.

In addition, we designed a simple iterative knowledge distillation method [7] (i.e., Teacher Network) to refine the node classification probabilities. The key idea is to train our proposed model (Student) to imitate the outputs of a pre-train a vanilla GNN (Teacher) by adding a regularization term of KL-divergence. This leads to solving the following optimization problem:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda_1 \text{KL}[\text{Dir}(\boldsymbol{\alpha}) \| \text{Dir}(\hat{\boldsymbol{\alpha}})] + \lambda_2 \text{KL}[P(\mathbf{y} \mid A, \mathbf{r}; \mathcal{G}) \parallel P(\mathbf{y}|\hat{\mathbf{p}})], \tag{11}$$

where $\hat{\mathbf{p}}$ is the vanilla GNN's (Teacher) output and λ_1 and λ_2 are trade-off parameters.

6 Experiments

In this section, we conduct experiments on the tasks of misclassification and OOD detections to answer the following questions for semi-supervised node classification:

Q1. Misclassification Detection: What type of uncertainty is the most promising indicator of high confidence in node classification prediction?

Q2. OOD Detection: What type of uncertainty is a key indicator to effectively detect OOD node?

Q3. GKDE with Uncertainty Estimates: How can GKDE help enhance prediction tasks with what types of uncertainty estimates?

Via extensive experiments, we found the following to answer the above questions:

A1. Dissonance (i.e., uncertainty due to conflicting evidence) played a significant role in increasing classification prediction accuracy as an indicator of confidence in an estimated prediction.

A2. Vacuity played a key role under our method S-BGCN-T-K particularly in OOD detection. Lack of information, leading to high unpredictability, helped detect OOD more effectively.

A3. GKDE can indeed help the accurate estimates of a Dirichlet distribution, leading to enhancing OOD detection particularly using vacuity.

6.1 Experiment Setup

Datasets: We use six datasets, including three citation network datasets [17] (i.e., Cora, Citeseer, Pubmed) and three new datasets [21] (i.e., Coauthor Physics, Amazon Computer, and Amazon Photo). We summarized the description and experimental setup of the used datasets in Appendix B.2[1].

Comparing Schemes: We conduct the extensive comparative performance analysis based on our proposed models and several state-of-the-art competitive counterparts. Here, all models are based on the most popular GNN model, GCN [12]. We compared our model (S-BGCN-T-K) against: (1) GCN [12] with the softmax probability entropy measuring uncertainty; and (2) Drop-GCN, adapting the Monte-Carlo Dropout [5, 16] into the GCN model to learn probabilistic uncertainty; (3) EDL-GCN, adapting the EDL model [18] with GCN to estimate evidential uncertainty; (4) DPN-GCN, adapting DPN [14] method with GCN to estimate probabilistic uncertainty. The performance is assessed by Area Under the ROC (AUROC) and Precision-Recall (AUPR) curves in both experiments [6].

6.2 Results

Misclassification Detection. The misclassification detection experiment involves detecting whether a given prediction is incorrect using an uncertainty measure. Table 1 shows that S-BGCN-T-K outperforms all baselines under AUROC and AUPR for misclassification detection. The outperformance of dissonance-based detection is fairly impressive. This confirms that low dissonance (a small amount of conflicting evidence) is the key to maximize the accuracy of node classification prediction. We observe the following performance order: $\texttt{Dissonance} > \texttt{Entropy} \approx \texttt{Aleatoric} > \texttt{Vacuity} \approx \texttt{Epistemic}$, which is aligned with our conjecture: higher dissonance with conflicting prediction leads to higher misclassification detection. We also conducted experiments on another three datasets, and obtained the similar trends of the results, as shown in Appendix C.

Table 1: AUROC and AUPR for the Misclassification Detection.

Data	Model	AUROC					AUPR					Acc
		Va.	Dis.	Al.	Ep.	En.	Va.	Dis.	Al.	Ep.	En.	
Cora	S-BGCN-T-K	70.6	**82.4**	75.3	68.8	77.7	90.3	**95.4**	92.4	87.8	93.4	**82.0**
	EDL-GCN	70.2	81.5	-	-	76.9	90.0	94.6	-	-	93.6	81.5
	DPN-GCN	-	-	78.3	75.5	77.3	-	-	92.4	92.0	92.4	80.8
	Drop-GCN	-	-	73.9	66.7	76.9	-	-	92.7	90.0	93.6	81.3
	GCN	-	-	-	-	79.6	-	-	-	-	94.1	81.5
Citeseer	S-BGCN-T-K	65.4	**74.0**	67.2	60.7	70.0	79.8	**85.6**	82.2	75.2	83.5	**71.0**
	EDL-GCN	64.9	73.6	-	-	69.6	79.2	84.6	-	-	82.9	70.2
	DPN-GCN	-	-	66.0	64.9	65.5	-	-	78.7	77.6	78.1	68.1
	Drop-GCN	-	-	66.4	60.8	69.8	-	-	82.3	77.8	83.7	70.9
	GCN	-	-	-	-	71.4	-	-	-	-	83.2	70.3
Pubmed	S-BGCN-T-K	64.1	**73.3**	69.3	64.2	70.7	85.6	**90.8**	88.8	86.1	89.2	**79.3**
	EDL-GCN	62.6	69.0	-	-	67.2	84.6	88.9	-	-	81.7	79.0
	DPN-GCN	-	-	72.7	69.2	72.5	-	-	87.8	86.8	87.7	77.1
	Drop-GCN	-	-	67.3	66.1	67.2	-	-	88.6	85.6	89.0	79.0
	GCN	-	-	-	-	68.5	-	-	-	-	89.2	79.0

Va.: Vacuity, Dis.: Dissonance, Al.: Aleatoric, Ep.: Epistemic, En.: Entropy

OOD Detection. This experiment involves detecting whether an input is out-of-distribution given a measure of uncertainty. For semi-supervised node classification, we randomly selected 1-4 categories as OOD categories and trained the models only based on training nodes of the other categories. Due to the space constraint, the experimental setup for the OOD detection is detailed in Appendix B.3.

In Table 2, across six network datasets, particularly our vacuity-based detection significantly outperformed among all, strikingly exceeding the performance of the epistemic uncertainty and other type of uncertainties. This proves that vacuity (lack of evidence) is the key to improve OOD detection performance. We observe the following performance order: $\texttt{Vacuity} > \texttt{Entropy} \approx \texttt{Aleatoric} > \texttt{Epistemic} \approx \texttt{Dissonance}$, which is consistent with the theoretical proof shown in Theorem 1.

Ablation Study. We conducted an additional experiments (see Table 3) in order to clearly demonstrate the contributions of the key technical components, including GKDE, Teacher Network, and subjective Bayesian framework. The key findings obtained from this experiment are: (1) GKDE can enhance

[1]The source code and datasets are accessible at https://github.com/zxj32/uncertainty-GNN

Table 2: AUROC and AUPR for the OOD Detection.

Data	Model	AUROC					AUPR				
		Va.	Dis.	Al.	Ep.	En.	Va.	Dis.	Al.	Ep.	En.
Cora	S-BGCN-T-K	**87.6**	75.5	85.5	70.8	84.8	**78.4**	49.0	75.3	44.5	73.1
	EDL-GCN	84.5	81.0		-	83.3	74.2	53.2	-	-	71.4
	DPN-GCN	-	-	77.3	78.9	78.3	-	-	58.5	62.8	63.0
	Drop-GCN	-	-	81.9	70.5	80.9	-	-	69.7	44.2	67.2
	GCN	-	-	-	-	80.7	-	-	-	-	66.9
Citeseer	S-BGCN-T-K	**84.8**	55.2	78.4	55.1	74.0	**86.8**	54.1	80.8	55.8	74.0
	EDL-GCN	78.4	59.4	-	-	69.1	79.8	57.3	-	-	69.0
	DPN-GCN	-	-	68.3	72.2	69.5	-	-	68.5	72.1	70.3
	Drop-GCN	-	-	72.3	61.4	70.6	-	-	73.5	60.8	70.0
	GCN	-	-	-	-	70.8	-	-	-	-	70.2
Pubmed	S-BGCN-T-K	**74.6**	67.9	71.8	59.2	72.2	**69.6**	52.9	63.6	44.0	56.5
	EDL-GCN	71.5	68.2	-	-	70.5	65.3	53.1	-	-	55.0
	DPN-GCN	-	-	63.5	63.7	63.5	-	-	50.7	53.9	51.1
	Drop-GCN	-	-	68.7	60.8	66.7	-	-	59.7	46.7	54.8
	GCN	-	-	-	-	68.3	-	-	-	-	55.3
Amazon Photo	S-BGCN-T-K	**93.4**	76.4	91.4	32.2	91.4	**94.8**	68.0	92.3	42.3	92.5
	EDL-GCN	63.4	78.1	-	-	79.2	66.2	74.8	-	-	81.2
	DPN-GCN	-	-	83.6	83.6	83.6	-	-	82.6	82.4	82.5
	Drop-GCN	-	-	84.5	58.7	84.3	-	-	87.0	57.7	86.9
	GCN	-	-	-	-	84.4	-	-	-	-	87.0
Amazon Computer	S-BGCN-T-K	**82.3**	76.6	80.9	55.4	80.9	**70.5**	52.8	60.9	35.9	60.6
	EDL-GCN	53.2	70.1	-	-	70.0	33.2	43.9	-	-	45.7
	DPN-GCN	-	-	77.6	77.7	77.7	-	-	50.8	51.2	51.0
	Drop-GCN	-	-	74.4	70.5	74.3	-	-	50.0	46.7	49.8
	GCN	-	-	-	-	74.0	-	-	-	-	48.7
Coauthor Physics	S-BGCN-T-K	**91.3**	87.6	89.7	61.8	89.8	**72.2**	56.6	68.1	25.9	67.9
	EDL-GCN	88.2	85.8	-	-	87.6	67.1	51.2	-	-	62.1
	DPN-GCN	-	-	85.5	85.6	85.5	-	-	59.8	60.2	59.8
	Drop-GCN	-	-	89.2	78.4	89.3	-	-	66.6	37.1	66.5
	GCN	-	-	-	-	89.1	-	-	-	-	64.0

Va.: Vacuity, Dis.: Dissonance, Al.: Aleatoric, Ep.: Epistemic, En.: Entropy

OOD detection (i.e., 30% increase with vacuity), which is consistent with our theoretical analysis of proving the outperformance of GKDE in uncertainty estimation, i.e., OOD nodes show high vacuity; and (2) The Teacher Network can further improve the node classification accuracy.

6.3 Why is Epistemic Uncertainty Less Effective than Vacuity?

Although epistemic uncertainty is known to be effective to improve OOD detection [5, 11] in computer vision applications, our results demonstrate it is less effective than our vacuity-based approach. The first potential reason is that epistemic always smaller than vacuity (From Theorem 1), which indicate that epistemic may capture less information related to OOD. Another potential reason is that the previous success of epistemic in computer vision applications are only applied in supervised learning, but they are not sufficiently validated in semi-supervised learning. Recall that epistemic uncertainty (i.e., model uncertainty) is calculated by mutual information (see Eq. (5)). In semi-supervised setting, the features of unlabeled nodes are also fed to a model for training process to provide the model with a high confidence in its output. For example, the small amount of mutual information due to the model output $P(\mathbf{y}|A, \mathbf{r}; \theta)$ is similar with a different sampled parameter θ. We also designed a semi-supervised learning experiment for image classification and observed a same pattern where the results are demonstrated in Appendix C.6.

Table 3: Ablation study of our proposed models: (1) `S-GCN`: Subjective GCN with vacuity and dissonance estimation; (2) `S-BGCN`: S-GCN with Bayesian framework; (3) `S-BGCN-T`: S-BGCN with a Teacher Network; (4) `S-BGCN-T-K`: S-BGCN-T with GKDE to improve uncertainty estimation.

Data	Model	AUROC (Misclassification Detection)					AUPR (Misclassification Detection)					Acc
		Va.	Dis.	Al.	Ep.	En.	Va.	Dis.	Al.	Ep.	En.	
Cora	S-BGCN-T-K	70.6	82.4	75.3	68.8	77.7	90.3	**95.4**	92.4	87.8	93.4	82.0
	S-BGCN-T	70.8	**82.5**	75.3	68.9	77.8	90.4	**95.4**	92.6	88.0	93.4	**82.2**
	S-BGCN	69.8	81.4	73.9	66.7	76.9	89.4	94.3	92.3	88.0	93.1	81.2
	S-GCN	70.2	81.5	-	-	76.9	90.0	94.6	-	-	93.6	81.5
		AUROC (OOD Detection)					AUPR (OOD Detection)					
Amazon Photo	S-BGCN-T-K	**93.4**	76.4	91.4	32.2	91.4	**94.8**	68.0	92.3	42.3	92.5	-
	S-BGCN-T	64.0	77.5	79.9	52.6	79.8	67.0	75.3	82.0	53.7	81.9	-
	S-BGCN	63.0	76.6	79.8	52.7	79.7	66.5	75.1	82.1	53.9	81.7	-
	S-GCN	64.0	77.1	-	-	79.6	67.0	74.9	-	-	81.6	-

Va.: Vacuity, Dis.: Dissonance, Al.: Aleatoric, Ep.: Epistemic, En.: Entropy

7 Conclusion

In this work, we proposed a multi-source uncertainty framework of GNNs for semi-supervised node classification. Our proposed framework provides an effective way of predicting node classification and detecting OOD considering multiple types of uncertainty. We leveraged the estimation of various types of uncertainty from both DL and evidence/belief theory domains. Via our extensive experiments, we found that dissonance-based detection yielded the best performance on misclassification detection while vacuity-based detection performed the best for OOD detection among all other competitive counterparts. In particular, it was noticeable that applying GKDE and the Teacher network further enhanced prediction accuracy along with accurate uncertainty estimation.

Acknowledgments

We would like to thank Yuzhe Ou for providing proof suggestions. This work is partially supported by ARL's Competitive Basic Research Program under Computational and Information Sciences Directorate and by the US Army Research Office under grant number W911NF1720129. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Broader Impact

In this paper, we propose a uncertainty-aware semi-supervised learning framework of GNN for predicting multi-dimensional uncertainties for the task of semi-supervised node classification. Our proposed framework can be applied to a wide range of applications, including computer vision, natural language processing, recommendation systems, traffic prediction, generative models and many more [24]. Our proposed framework can be applied to predict multiple uncertainties of different roots for GNNs in these applications, improving understanding of individual decisions, as well as the underlying models. While there will be important impacts resulting from the use of GNNs in general, here we focus on the impact of using our method to predict multi-source uncertainties for such systems. There are many benefits to using our method, such as improvement of safety and transparency in decision-critical applications to avoid overconfidence prediction.

We see opportunities for research applying our uncertainty framework to beneficial purposes, such as investigating whether this uncertainty framework could improve misclassification detection or OOD detection. To mitigate the risks associated with using multi-source uncertainty, we encourage research to understand the impacts of using uncertainty framework in particular real-world scenarios.

References

[1] C. W. De Silva. *Intelligent control: fuzzy logic applications*. CRC press, 2018.

[2] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv preprint arXiv:1710.07283*, 2017.

[3] D. Eswaran, S. Günnemann, and C. Faloutsos. The power of certainty: A dirichlet-multinomial model for belief propagation. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2017.

[4] Y. Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

[5] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.

[6] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[8] A. Jøsang. *Subjective logic*. Springer, 2016.

[9] A. Jøsang, J.-H. Cho, and F. Chen. Uncertainty characteristics of subjective opinions. In *FUSION*, pages 1998–2005. IEEE, 2018.

[10] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[11] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] Z.-Y. Liu, S.-Y. Li, S. Chen, Y. Hu, and S.-J. Huang. Uncertainty aware graph gaussian process for semi-supervised learning. 2020.

[14] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

[15] N. J. Nilsson. Probabilistic logic. *Artificial intelligence*, 28(1):71–87, 1986.

[16] S. Ryu, Y. Kwon, and W. Y. Kim. Uncertainty quantification of molecular property prediction with bayesian neural networks. *arXiv preprint arXiv:1903.08375*, 2019.

[17] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[18] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *NIPS*, pages 3183–3193, 2018.

[19] K. Sentz, S. Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Citeseer, 2002.

[20] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[21] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

[22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks. *ICLR*, 2018.

[23] Y. Zhang, S. Pal, M. Coates, and D. Üstebay. Bayesian graph convolutional neural networks for semi-supervised classification. *arXiv preprint arXiv:1811.11103*, 2018.

[24] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.