A Proof of Proposition 2.2: additive expansion proposition

We denote the embedding vector of a node v_i by $e_i \triangleq g_i(\hat{G}) = \text{GNN}(\hat{G})[i]$. Without loss of generality, we drop the subscript for short. We can also define the density measure d_f as:

$$p_f(e) \triangleq \frac{\exp(-d(e))}{\int_{\mathcal{G}} \exp(-d(e))dg(\mathcal{G})}. \tag{5}$$

For any subset of the embedding space $E \subset g(\mathcal{X})$, the local probability can be measured by $p_f(E) = \int_E p_f(e)de$. If we have a local optimal subset $U \subset E$ with a confidence threshold of $1-q$, and its perturbation U_ϵ, then the consistency at the boundary of the subset separation problem $E \to U, E \setminus U$ can be quantified by Cheeger constant. We introduce the continuous Cheeger inequality to elaborate the lower bound of the Cheeger constant under an amplified measure αf, where $\alpha > 1$ is a constant.

We first define the restricted *Cheeger constant* in the link prediction task. Given the function f and any subset E, Cheeger constant is calculated by

$$\mathcal{C}_f(E) \triangleq \lim_{\epsilon \to 0^+} \inf_{A \subset E} \frac{p_f(A_\epsilon) - p_f(A)}{\epsilon \min\{p_f(A), p_f(E \setminus A)\}}. \tag{6}$$

According to the definition, the Cheeger constant is a lower bound of probability density in the neighborhood of the given set. It quantifies the chance of escaping the subset A under the probability measure f and reveals the consistency over the set cutting boundary.

Then we prove that for the any subset $E \subset g(\mathcal{G})$ with its local optimal subset $U : \{e \in E : p_f(e) > 1-q\}$, there exists $\alpha > 1$ s.t. $\mathcal{C}_{\alpha f}(E \setminus U) \geq 1$.

As the measurable function for the link prediction is defined as $f(e) = -e^T e_a$. When $e^* = e_a$, f reaches the global minimal. For the embedding vectors outside the local minimal subset $e_y \in g(\mathcal{G}) \backslash U$, there exsits $\epsilon > 0$ s.t.

$$f(e_y) \geq f(e^*) + 2\hat{\epsilon}, \tag{7}$$

where $\hat{\epsilon} = C\epsilon$. If we define $E^*_\epsilon = \{e^*_{\hat{\epsilon}}\} \cap g(\mathcal{G})$, where $e^*_{\hat{\epsilon}}$ is the $\hat{\epsilon}$ neighbor of e^*, according to the Lipchitz condition of f, for $e \in E^*_\epsilon$, we have:

$$f(e_x) \leq f(e^*) + \hat{\epsilon}\|e_x - e^*\|_2 \leq f(e^*) + \hat{\epsilon}. \tag{8}$$

Combining Eq.8 and Eq.7 leads to $f(e_y) - f(e_x) \geq \hat{\epsilon}$. Thus, for the amplified probability measure $p_{\alpha f}$, we have

$$p_{\alpha f}(e_x)/p_{\alpha f}(e_y) \geq \exp(\alpha\hat{\epsilon}) \tag{9}$$

According to the inequality property from [26] (formula 63), we have

$$\frac{p_{\alpha f}(U)}{p_{\alpha f}(g(\mathcal{G}) \setminus U)} \geq \exp\left(\alpha\hat{\epsilon} - 2\text{log}\left(2C^2/\hat{\epsilon}\right)\right). \tag{10}$$

As $p_{\alpha f}(g(\mathcal{G}) \setminus U) + p_{\alpha f}(U) = 1$. If we select α large enough s.t. the RHS of Eq.10 is larger than 1, $p_{\alpha f}(g(\mathcal{G}) \setminus U) \leq \frac{1}{2}$. Thus, according to [17] (Theorem 2.6), we have $\mathcal{C}_{\alpha f} \geq 1$. It guarantees consistency around the perimeter of the U. As $\alpha > 1$ and $p_f, p_{\alpha f}$ are bounded on any subsets of embedding space. It implies a probability margin $\eta > 0$ at the neighborhood of the local optimal between two measurable functions $f, \alpha f$, where

$$\eta = \inf_{\hat{e} \in U_\epsilon \backslash U, e \in U} (p_{\alpha f}(\hat{e}) - p_f(e)). \tag{11}$$

which, according to [23], implies additive expansion property of the probability measure in the link prediction, as Proposition 2.2.

B Proof of Theorem 2.3: error analysis

In [23], $\mathcal{M}(g_\phi)$ is also assumed to satisfy additive-expansion (q, ϵ), where $\mathcal{M}(g) \triangleq \{y \in Y : g(y) \neq y\}$ is the set of mis-classified samples, and they give the error bound of the trained classifier s (Theorem B.2):

$$\mathrm{Err}(g) \leq 2(q + \mathcal{A}(g)). \tag{12}$$

Here in link prediction task, $\mathcal{M}(g_\phi)$ is mis-classified samples by the pseudo labeler (teacher model). It can be written by $\{y_i : Y_p[i] \neq \mathcal{E}_\mathcal{T}[i]\}$, which is intractable during the training. The probability threshold is $1 - q$ and a local optimal subset U for PL is constructed accordingly. We aim to let $\mathcal{M}(g_\phi) \cap U$ be close to $\emptyset$, so that $g(\mathcal{G}) \setminus U$ can cover $\mathcal{M}(g_\phi)$ as much as possible. So we define the robust set $\mathcal{S}(g)$ as

$$\mathcal{S}(g) = \{y : g\left(y\right) = g\left(\hat{y}\right), \hat{y} \in \{y_\epsilon\}\}, \tag{13}$$

where y_ϵ is the ϵ neighborhood of sample y. Then, according to Proposition 2.1, we have:

$$p_f\left(\{y \in Y : g_\phi(y) \neq y, y \in \mathcal{S}(g_\psi)\}\right) \leq p_{\alpha f}\left(g\left(\mathcal{G}\right) \setminus U\right) \leq q, \tag{14}$$

which has similar form with [23] Lemma B.3 for link prediction task. Besides, the analysis of $p_f(\{y \in Y : g_\phi(y) = y, g_\psi(y) \neq y, y \in \mathcal{S}(g_\psi)\})$ and $p_f(\overline{\mathcal{S}(g_\psi)})$ are the same. Thus, the assumption on $\mathcal{M}(g_\phi)$ is satisfied. Then, we can draw the same conclusion with Eq.12, and the classifier is the student model g_ψ. The theorem is proofed.

C Proof of convergence inequality

The PL strategy $\mathcal{T}$ for the unlabeled data provides a Bayesian prior, from which we formalize the empirical loss defined in Eq.1 as

$$\mathcal{L}_\mathcal{T}^{(t+1)} = \frac{1}{\left|\hat{Y}_o^{(t)}\right| + k}\left[\mathrm{CE}\left(g_\psi^{(t)}, \hat{Y}_o^{(t)}\right) + \mathrm{CE}\left(g_\psi^{(t)}, Y_p^{(t)}\right)\right]. \tag{15}$$

We can decompose the cross-entropy loss of the pseudo labeled samples by:

$$\begin{aligned}\mathrm{CE}\left(g_\psi, Y_p\right) =& \sum_{\hat{Y}_u} \mathrm{ce}\left(g_\psi, Y\right) \cdot \mathcal{T} \\ =& \sum_{\hat{Y}_u} \left[\mathrm{ce}\left(g_\psi, Y\right) - Y\left[\mathrm{ce}\left(g_\psi, Y\right)\right]\right] \cdot \left[\mathcal{T} - Y\mathcal{T}\right] \\ &+ Y\mathcal{T}\sum_{\hat{Y}_u} \mathrm{ce}\left(g_\psi, Y\right) + \mathbb{E}\left[\mathrm{ce}\left(g_\psi, Y\right)\right]\sum_{\hat{Y}_u}\mathcal{T} - \left|\hat{Y}_u\right| Y\mathcal{T}Y\left[\mathrm{ce}\left(g_\psi, Y\right)\right]\end{aligned} \tag{16}$$

.

Thus, Eq.16 can be simplified to:

$$\begin{aligned}\mathrm{CE}\left(g_\psi, Y_p\right) =& \left|\hat{Y}_u\right| \mathrm{Cov}\left[\mathrm{ce}\left(g_\psi, Y\right), \mathcal{T}\right] + \mathbb{E}\mathcal{T} \cdot \left|\hat{Y}_u\right| \mathbb{E}\left[\mathrm{ce}\left(g_\psi, Y\right)\right] \\ &+ \mathbb{E}\left[\mathrm{ce}\left(g_\psi, Y\right)\right] \cdot \left|\hat{Y}_u\right| \mathbb{E}\mathcal{T} - \left|\hat{Y}_u\right| \mathbb{E}\mathcal{T}\mathbb{E}\left[\mathrm{ce}\left(g_\psi, Y\right)\right] \\ =& \left|\hat{Y}_u\right| \mathrm{Cov}\left[\mathrm{ce}\left(g_\psi, Y\right), \mathcal{T}\right] + \left|\hat{Y}_u\right| \mathbb{E}\mathcal{T}\mathbb{E}\left[\mathrm{ce}\left(g_\phi, Y\right)\right] \\ =& \left|\hat{Y}_u\right| \mathrm{Cov}\left[\mathrm{ce}\left(g_\psi, Y\right), \mathcal{T}\right] + k\mathbb{E}\left[\mathrm{ce}\left(g_\phi, Y\right)\right]\end{aligned} \tag{17}$$

Note that $\mathcal{T}$ is the indicator-like function, where we have

$$\mathbb{E}\mathcal{T} = \frac{1}{\left|\hat{Y}_u\right|}\sum_{\hat{Y}_u}\mathcal{T} = \frac{k}{\left|\hat{Y}_u\right|}. \tag{18}$$

Table 6: Details of Node Information in the Case Study.

Node	Group 1			Group 2	
	Node 2702	**Node 5688**	**Node 8906**	**Node 3489**	**Node 7680**
ID	17505908	11353631	30138652	23221074	12265137
Outlinks	[6097297]	[6097297]	[244374, 6097297]	[20901]	[]
Title	Ubuntu Hacks	Pungi (software)	LinuxPAE64	Malware Bell	Norton Confidential
Label	Operating systems	Operating systems	Operating systems	Computer security	Computer security
Tokens	"ubuntu", "hacks", "ubuntu", "hacks", "tips", "tools", "exploring", "using", "tuning", "linux", "book", "tips", "ubuntu", "popular", "linux", "distribution", "book", "published", "o'reilly", "media", "june", "2006", "part", "o'reilly", "hacks", "series"	"pungi", "software", "pungi", "program", "making", "spins", "fedora", "linux","distribution", "release", "7", "upwards"	"linuxpae64", "linuxpae64", "port", "linux", "kernel", "running", "compatibility", "mode", "x86-64", "processor", "kernel", "capable", "loading", "i386", "modules", "device", "drivers", "supports", "64-bit", "linux", "applications", "user", "mode"	"malware", "bell", "malware", "bell", "malware", "program", "made", "taiwan", "somewhere", "2006", "2007", "malware", "bell", "tries", "install", "automatically", "upon", "visiting", "website", "promoting", "containing", "malware"	"norton", "confidential", "norton", "confidential", "program", "designed", "encrypt", "passwords", "online", "detect", "phishing", "sites"

Based on the Eq.15 and Eq.17, we can rewrite $\mathcal{L}_{\mathcal{T}}^{(t+1)}$ as

$$\begin{aligned} \mathcal{L}_{\mathcal{T}}^{(t+1)} &= \beta \text{Cov}\left[\text{ce}\left(g_{\psi}, Y\right), \mathcal{T}\right] + \frac{1}{\left|\hat{Y}_o^{(t)}\right|}\text{CE}\left(g_{\psi}^{(t)}, \hat{Y}_o^{(t)}\right) \\ &\leq \beta \text{Cov}\left[\text{ce}\left(g_{\psi}, Y\right), \mathcal{T}\right] + \frac{1}{\left|\hat{Y}_o^{(t)}\right|}\text{CE}\left(g_{\phi}^{(t)}, \hat{Y}_o^{(t)}\right) \\ &= \beta \text{Cov}\left[\text{ce}\left(g_{\psi}, Y\right), \mathcal{T}\right] + \mathcal{L}_{\mathcal{T}}^{(t)} \end{aligned} \tag{19}$$

where $\beta = |\hat{Y}_u|/(|\hat{Y}_o| + k$. The inequality holds due to the assumption.

D Case study of CPL on link prediction

Error bound: In the case study, the recorded confidence threshold is $1 - q = 0.98$ for WikiCS. We adopt 5 views of dropout with the augmentation drop rate 0.05. And according to the error bound given by Theorem 2.3, given the confidence threshold, Eq.2 suggests that the higher prediction consistency should lead to a smaller error bound. The final prediction consistency is $\mathcal{A}(g) = 0.0358$, thus, we can calculate error bound $Err(g) = 0.1116$. The AUC and AP are $95.56 \pm 0.24\%, 95.58 \pm 0.29\%$ which are bounded within $Err(g)$.

Knowledge discovery: In the 5 random experiments, we add 500 pseudo links in each iteration. Here we focus on the common PL links in the first iteration, which are considered the most confident samples. We look for the metadata of WikiCS whose node, feature, link and node label represent paper, token, reference relation and topic of the paper respectively. There are These 7 most confident links categorized into 2 groups. We take 3 out of 5 nodes in group1 and the 2 nodes in group2 for analysis, whose detailed information of these nodes is shown in AppendixD.

For group1, 3 nodes are connected by the pseudo links, and they are all linked to a central node whose degree is 321. The metadata information of the nodes are all strongly relevant to "Linux" in the "operating systems" topic. Thus, the PL linked nodes are likely to have common neighbors discovered triangle relationship. In group2, node 3489 has no in/out degree and is pseudo linked to node 7680. Both papers focus on the "malware"/"phishing" under the topic "Computer security". Although they only have one common token, the CPL strategy successfully discovers the correlation and consistently add it to the training set. The detailed result of the case study is shown in Table 6.