Approximate Inference in Structured Instances with Noisy Categorical Observations – Supplementary Material

Alireza Heidari, Ihab F. Ilyas, Theodoros Rekatsinas

1 ANALYSIS FOR TREES

1.1 Proof of Lemma 1

Proof. In $G = (V, E)$, for each edge $(u, v) \in E$, we have a random variable $L_{u,v} = \mathbb{1}(\varphi(Z_u, Z_v) \neq X_{u,v})$ with distribution:

$$L_{u,v} = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$$

To apply the Bernstein inequality, we must consider $L_{u,v} - p$. We have $E[L_{u,v} - p] = 0$ and $\sigma^2(L_{u,v} - p) = p(1-p)$. We must also have that the random variables are constrained. We know that $|L_{u,v} - p| \leq \max\{1-p, p\}$ and $p < 1/2$ so $|L_{u,v} - p| \leq 1 - p$. Now, we apply the Bernstein inequality:

$$P\left(\sum_{(u,v)\in E} L_{u,v} - p \leq t\right) \geq 1 - \exp\left(-\frac{t^2}{2|E|\sigma^2 + \frac{2}{3}(1-p)t}\right)$$

Let $u \triangleq -\frac{t^2}{2|E|\sigma^2 + \frac{2}{3}t(1-p)}$. Solving for t we obtain:

$$t = \frac{1}{3}u(1-p) + \sqrt{\frac{(1-p)^2u^2}{9} + 2|E|\sigma^2 u}$$

Now we have that:

$$P\left(\sum L_{u,v} - p \leq \frac{1}{3}u(1-p) + \sqrt{\frac{(1-p)^2u^2}{9} + 2|E|\sigma^2 u}\right) \geq 1 - e^{-u}$$

We choose $u = \ln\left(\frac{2}{\delta}\right)$, and substituting $|E| = n - 1$ for trees and $\sigma^2 = p(1-p)$, we have that with probability $1 - \delta$:

$$\sum_{(u,v)\in E} \mathbb{1}(\varphi(u,v) \neq X_{u,v}) \leq \frac{1}{3}\ln(\frac{2}{\delta})(1-p) + \sqrt{\frac{(1-p)^2\ln(\frac{2}{\delta})^2}{9} + 2(n-1)p(1-p)\ln(\frac{2}{\delta})} + (n-1)p$$

Simplifying by noting that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have proven the lemma.

□

1.2 Proof of Lemma 3

Proof. For $\hat{Y}_v = Y_v$, we have that $P_Z(Y_v \neq Z_v) - P_Z(Y_v \neq Z_v) = 0$ and so we are done. When $\hat{Y}_v \neq Y_v$, we have that $P_Z(Y_v \neq Z_v) = q$ and get the following for the first term:

$$\begin{aligned} P_Z(\hat{Y}_v \neq Z_v) &= P_Z(\hat{Y}_v \neq Z_v \wedge Z_v = Y_v) + P_Z(\hat{Y}_v \neq Z_v \wedge Z_v \neq Y_v) \\ &= P_Z(\hat{Y}_v \neq Z_v \wedge Z_v = Y_v) + \sum_{i\in[k]\wedge i\neq\hat{Y}_v\wedge i\neq Y_v} P_Z(\hat{Y}_v \neq Z_v \wedge Z_v \neq Y_v \wedge Z_v = i) \end{aligned}$$

We know that $P_Z(\hat{Y}_v \neq Z_v \wedge Z_v = Y_v) = 1 - q$. For each i we have $P_Z(\hat{Y}_v \neq Z_v \wedge Z_v \neq Y_v \wedge Z_v = i) = \frac{q}{k-1}$. So we have:

$$\sum_{i \in [k] \wedge i \neq \hat{Y}_v \wedge i \neq Y_v} P_Z(\hat{Y}_v \neq Z_v \wedge Z_v \neq Y_v \wedge Z_v = i) = \frac{q}{k-1} \sum_{i \in [k] \wedge i \neq \hat{Y}_v \wedge i \neq Y_v} 1 = \frac{q(k-2)}{k-1}$$

Given this we have:

$$\begin{aligned} P_Z(\hat{Y}_v \neq Z_v) &= P_Z(\hat{Y}_v \neq Z_v \wedge Z_v = Y_v) + \sum_{i \in [k] \wedge i \neq \hat{Y}_v \wedge i \neq Y_v} P_Z(\hat{Y}_v \neq Z_v \wedge Z_v \neq Y_v \wedge Z_v = i) \\ &= (1-q) + \frac{q(k-2)}{k-1} = 1 - \frac{q}{k-1} \end{aligned}$$

Finally, consolidating these, we get:

$$P(\hat{Y}_v \neq Z_v) - P(Y_v \neq Z_v) = 1 - \frac{q}{k-1} - q = 1 - \frac{k}{k-1} q$$

This is exactly c, and so the hamming error and excess risk are proportional. Furthermore, we can set c to $1 - \frac{k}{k-1}q$. □

1.3 Proof of Corollary 1

Proof. Before the actual proof, we show that:

Lemma 1. *In trees, if $Y \in \mathcal{F}$ then $Y^* = Y$.*

Proof. We have that $Y \in \mathcal{F}$ hence $\sum_{(u,v)\in E} \mathbb{1}\{\varphi(Y_u, Y_v) \neq X_{u,v}\} \leq t$ with $t = (n-1)p + \frac{2}{3}\ln(2/\delta)(1-p) + \sqrt{2(n-1)p(1-p)\ln(2/\delta)}$ and it holds with probability $1-\delta$. We have that $Y^* = \arg\min_{Y' \in \mathcal{F}} \sum_{v \in V} \mathbb{P}(Y'_v \neq Y_v)$ which among all possible $Y' \in \mathcal{F}$ finds the one nearest to Y. So we have $Y^* = Y$. □

Let $Y^* = \arg\min_{\mathcal{Y} \in \mathcal{F}} \sum_{v \in V} \mathbb{P}(\mathcal{Y}_v \neq Z_v)$ and let $\hat{Y}$ be the ERM. Because $Y \in \mathcal{F}$ so $Y^* = Y$.Then from Lemma 2, we have:

$$\sum_{v \in V} P(\hat{Y}_v \neq Z_v) - \sum_{v \in V} P(Y^*_v \neq Z_v) \leq \left(\frac{2}{3} + \frac{c}{2}\right) \log\left(\frac{|\mathcal{F}|}{\delta}\right) + \frac{1}{c} \sum_{v \in V} \mathbb{1}\{\hat{Y}_v \neq Y_v\}$$

For all $c > 0$, we can use Lemma 3 and apply it to the RHS to obtain

$$\left(1 - \frac{1}{ct}\right) \left[\sum_{v \in V} P(\hat{Y}_v \neq Z_v) - \sum_{v \in V} P(Y^*_v \neq Z_v)\right] \leq \left(\frac{2}{3} + \frac{c}{2}\right) \log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

where $t = 1 - \frac{k}{k-1}q$. Now, because this holds for $c > 0$, we can choose $c = \frac{2}{t}$ thus obtaining

$$\left(\frac{1}{2}\right) \sum_{v \in V} P(\hat{Y}_v \neq Z_v) - \sum_{v \in V} P(Y^*_v \neq Z_v) \leq \left(\frac{2}{3} + \frac{1}{t}\right) \log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

Finally, applying that $q = \frac{1}{2} - \varepsilon$, we obtain our result. □

1.4 Proof of Theorem 1

Proof. By Lemma 1, we have with probability at least $1 - \frac{\delta}{2}$

$$\sum_{(u,v) \in E} \mathbb{1}(\varphi(u,v) \neq X_{u,v}) \leq t$$

which we will use it to define a hypothesis class $\mathcal{F}$ as

$$\mathcal{F} = \left\{ \hat{Y} : \sum_{(u,v)\in E} \mathbb{1}(\varphi(\hat{Y}_u, \hat{Y}_v) \neq X_{u,v}) \leq t \right\}$$

with

$$t = (n-1)p + \frac{2}{3}\ln(\frac{2}{\delta})(1-p) + \sqrt{2(n-1)p(1-p)\ln(\frac{2}{\delta})}$$

Which suggests that $Y \in \mathcal{F}$ with high probability. By Corollary 1, we have that $\hat{Y}$ being the ERM over $\mathcal{F}$ implies that

$$\sum_{v\in V} P(\hat{Y}_v \neq Z_v) - \min_{Y\in\mathcal{F}} \sum_{v\in V} P(Y_v \neq Z_v) \leq \left(\frac{4}{3} + \frac{2}{\frac{1}{4} + (\frac{1}{4} - \epsilon)(1 - \frac{k}{k-1})} \right) \log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

Combining this with Lemma 3, we conclude that $\sum_{v\in V} \mathbb{1}\{\hat{Y}_v \neq Y_v\}$ is bounded form above by

$$\frac{1}{1 - \frac{k}{k-1}q} \left(\frac{4}{3} + \frac{2}{\frac{1}{4} + (\frac{1}{4} - \epsilon)(1 - \frac{k}{k-1})} \right) \log\left(\frac{|\mathcal{F}|}{\delta}\right)$$

Now, we approximate the size of the class $\mathcal{F}$. We can do so by upper-bounding the number of ways to violate the observed measurements. Pessimistically, of the possible $l = 0, 1, \ldots t$ violations, there are at most l nodes which are involved in this violation. Furthermore, there are at most $k-1$ ways for each of these nodes to be involved in such a violation. Therefore, we have,setting $t = \frac{2}{3}\ln(2/\delta)(1-p) + \sqrt{2(n-1)p(1-p)\ln(2/\delta)} + (n-1)p$

$$|\mathcal{F}| \leq \sum_{l=0}^{t} \binom{n}{l} k^l \leq k^t \sum_{l=0}^{t} \binom{n}{l} \leq k^t 2^t$$

Using this bound for $|\mathcal{F}|$, and assuming that the noise and sampling distribution is constant, we obtain that the hamming error is bounded by $\tilde{O}(\log(k)np)$ □

1.5 Solving the Optimization Problem on Trees with Dynamic Programming

Because G is assumed to be a tree, we can compute optimal solutions to subproblems. Specifically, we can turn any undirected tree into a controlled one by a breadth-first search.

Then we can define a table $OPT(u, B|\ell)$ which stores optimal values to the subtree rooted at u, constrained to budget B and with the parent of u constrained to class ℓ. Given the values of OPT for all descendants of a node u, it is not difficult to find values for the table at u. We formalize this in the following theorem.

Theorem 3. *The optimization problem 1 can be solved in time $O(kn^3p)$.*

Proof. Given a tree $T = (V, E)$, a budget t, observations $X = \{X_{u,v}\}_{(u,v)\in E}$ and $Z = \{Z_v\}_{v\in V}$, we would like to compute a solution to

$$\sum_{(u,v)\in E} \mathbb{1}(\varphi(Z_u, Z_v) \neq X_{u,v}) \leq \frac{2}{3}\ln(2/\delta)(1-p) + \sqrt{2(n-1)p(1-p)\ln(2/\delta)} + (n-1)p$$

First, we turn T into a tree rooted at some node r by running a breadth-first search from r and directing nodes according to their time of discovery. Call this directed tree rooted at r $\overrightarrow{T}_r$. We specify a table OPT which will collect values of optimal subproblems.

Specifically, denote $\overrightarrow{T}_u$ as the subtree of $\overrightarrow{T}_r$ rooted at a node u. Then OPT will be a matrix parameterized by $OPT(u, B|\ell)$ where $u \in V$, $0 \leq B \leq |\overrightarrow{T}_u|$ (no tree can violate the observations more times than the number of nodes in the tree) and $\ell \in [k]$. Let $Pa(u)$ be the singular parent of the node u. Then OPT values represent the optimal value of the subtree rooted at u with a budget B and $Pa(u)$ restricted to the value ℓ. Our recursive equation for OPT is then

$$OPT(u, B|i) = \min_{\ell\in[k]} \min_{\substack{\sum\limits_{v\in N(u)} B_v \\ = B - \mathbb{1}\{X_{u,v}\neq\varphi(i,\ell)\}}} \sum_{v\in N(u)} OPT(v, B_v|\ell) + \mathbb{1}\{\ell \neq Z_u\}$$

If we have the value of $OPT(u, B|\ell)$ for all nodes $u \neq r$, values ℓ and valid budgets $B \leq t$, we can calculate the optimum value of the tree by the following: We attach a node r' to r by an edge $r' \to r$ and set the information on the node to $X_{r',r} = 1$ then solve $OPT(r, t|1)$, then repeat the process but with $X_{r',r} = -1$, return the smaller of these two values.

For a leaf node w, the value of $OPT(w, B'|\ell)$ is simply $\min_i \mathbb{1}\{i \neq Z_w\}$ for $B' = 1$. If $B' = 0$ then we must choose i such that it does not violate the side information, i.e. we must have $\varphi(i, \ell) = X_{w,Pa(w)}$

Finally we show how to compute the summation in (**??**) efficiently. For each value $\ell \in [k]$ we must optimize the summation

$$\min_{\substack{\sum\limits_{v\in N(u)} B_v \\ = B - \mathbb{1}\{X_{u,v}\neq\varphi(i,\ell)\}}} \left(\sum_{v\in N(u)} OPT(v, B_v|\ell) + \mathbb{1}\{\ell \neq Z_u\} \right)$$

Because each node's optimal value is independent, we can rewrite this sum by submitting an optional order on $N(u)$ of $1, 2, \ldots, m = |N(u)|$ and reforming this sum to

$$\min_{B_1\in[0,K-\mathbb{1}\{\varphi(\ell,s)\neq X_{u,Pa(u)}\}]} OPT(1, B_1|\ell) + \min_{\sum_{j\in[2,m]} B_j = B - B_1 - \mathbb{1}\{\varphi(\ell,s)\neq X_{u,Pa(u)}\}} \sum_{j\in 2,m} OPT(j, B_j|\ell)$$

The minimization for the first two vertices whose number of constraints violated are at most B can be solved in $O(B^2)$ time. The calculation for the first three vertices can then be done in $O(B^2)$ time by reusing the information from the first two. We can repeat this until we have considered all children of u. Hence because we must calculate this value for all k possible classes, we get an algorithm which takes time $k \sum_{v\in V} |N(v)|B^2 = O(nkB^2)$. The statistical analysis below shows that B is *poly*(n, p). □

2 ANALYSIS FOR GENERAL GRAPHS

2.1 Approximation Correlation Clustering

We have following Theorem,

Theorem 4. *(Giotis & Guruswami, 2006) There is a polynomial time factor* 0.878 *approximation algorithm for* MaxAgree[2] *on general graphs. For every* $k \geq 3$, *there is a polynomial time factor* 0.7666 *approximation algorithm for* MaxAgree[k] *on general graphs.*

With this assumption in the worse case, we have labeling with 0.7666Opt[k]. If Opt $= |E| - b$ which b is the number of bad edges that the optimal does not cover. We know the original graph is a k cluster with no bad-cycle (a cycle with one negative edge), so whatever bad edges that we see are the result of the noise process on the edges, so $b \leq |E|p$ because part of them do not generate bad-cycles. We can consider the approximate process as an extra source to generate more bad edges so we have $|E| - b' \geq$ Approx$[k] = 0.7666$Opt$[k]$. Also, by our assumption we have $p \leq p'$ so $b \leq b'$

$$|E| - b' \geq \textsc{Approx}[k] = 0.7666\textsc{Opt}[k] = 0.7666\big(|E| - b\big) \to b' \leq 0.2334|E| + 0.7666b$$

So we have

$$b \leq b' \leq 0.2334|E| + 0.7666b$$

We have upper bound for the error introduced by our approximation and we assume all that noise come from edge noise process and the correlation clustering could not correct it, we can assume a noise process with p' such that $b' = |E|p'$ so :

$$|E|p' = b' \leq 0.2334|E| + 0.7666b \leq 0.2334|E| + 0.7666|E|p \rightarrow p' \leq 0.2334 + 0.7666p$$

So we consider exact correlation clustering result in our analyses and if we interested to see the effect of approximation algorithm on the result and get an error bound, we update p to $0.2334 + 0.7666p$ as worst case analysis which means we directly inject the approximation noise error to the results. This assumption is weak because part of b' can be captured by the local and global optimizer which we neglect it.

2.2 Proof of Lemma 4

Proof. We mix two partitions into one notation and each data point in D shows as $v_i = (\bar{Y}_i, Z_i)$, for each $i \in D$, and $\bar{Y}_i \in \bar{Y}$ and $Z_i \in Z$. We define $\forall l \in [k]$

$$X_l = \{v_i | \bar{Y}_i = l\}$$
$$T_l = \{v_i | Z_i = l\}$$

and the error is $E = \sum\limits_{v_i \in D} \mathbb{1}\{Z_i \neq \bar{Y}_i\}$. The only thing that we allowed to change is the label of X_ls. We can represent the partition X and T as,

$$X = \{X_1, X_2, \dots, X_k\}$$
$$T = \{T_1, T_2, \dots, T_k\}$$

We claim that with Algorithm 2, we can find the permutation π on X, such that E minimize. Let π^* be the permutation that makes minimum E. We prove this theorem with reductio ad absurdum. Therefore

$$E_{\pi^*} \leq E_{\pi} \tag{1}$$

Let N be the set of all $v_i \in D$ such that $\pi(\bar{Y}_i) \neq \pi^*(\bar{Y}_i)$,

$$N = \{v_i \in D | \pi(\bar{Y}_i) \neq \pi^*(\bar{Y}_i)\}$$

We can write E for π,

$$\begin{aligned} E_{\pi} &= \sum_{v_i \in D} \mathbb{1}\{\pi(\bar{Y}_i) \neq Z_i\} \\ &= \sum_{v_i \in N} \mathbb{1}\{\pi(\bar{Y}_i) \neq Z_i\} + \sum_{v_i \notin N} \mathbb{1}\{\pi(\bar{Y}_i) \neq Z_i\} \end{aligned}$$

Similarly we can define E_{π^*},

$$\begin{aligned} E_{\pi^*} &= \sum_{v_i \in D} \mathbb{1}\{\pi^*(\bar{Y}_i) \neq Z_i\} \\ &= \sum_{v_i \in N} \mathbb{1}\{\pi^*(\bar{Y}_i) \neq Z_i\} + \sum_{v_i \notin N} \mathbb{1}\{\pi^*(\bar{Y}_i) \neq Z_i\} \end{aligned}$$

Second term in E_{π^*} and E_{π} are equal, using Inequality 1, and we define $E_{\pi}(N) = \sum\limits_{v_i \in N} \mathbb{1}\{\pi(\bar{Y}_i) \neq Z_i\}$ and similarly $E_{\pi^*}(N)$ for π^*, so we have,

$$E_{\pi^*}(N) \leq E_{\pi}(N) \tag{2}$$

We know $N \subseteq D$, so the partition X on D present a sub-partition $\hat{X}$ on N. $\hat{X}$ defines like X, so $\hat{X} = \{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k\}$. This sub-partition notion can be defined for both permutations π and π^*,

$$\hat{X}_\pi = \{\hat{X}_{\pi(1)}, \hat{X}_{\pi(2)}, \ldots, \hat{X}_{\pi(k)}\}$$
$$\hat{X}_{\pi^*} = \{\hat{X}_{\pi^*(1)}, \hat{X}_{\pi^*(2)}, \ldots, \hat{X}_{\pi^*(k)}\}$$

In the greedy algorithm, we sort the intersections of X_is and T_is and select the biggest one each time, because $\hat{X}$ is sub-partition of X, so we have,

$$\forall v_i, v_j \in \hat{X} \quad \pi(\bar{Y}_i) = \pi(\bar{Y}_j) \longleftrightarrow \pi^*(\bar{Y}_i) = \pi^*(\bar{Y}_j) \tag{3}$$

Based on Equation 3, we can define a isomorphism on N,

$$\forall v_i \in N \quad \phi : \pi^*(\bar{Y}_i) \to \pi(\bar{Y}_i)$$

we define $\mathrm{m\dot{a}x}()$ as selecting the set with maximum size among all feasible sets, then we have,

$$\hat{X}_{\pi(\bar{Y}_i)} = \left\{ v_j \in D | \pi(\bar{Y}_i) = \pi(\bar{Y}_j), \pi(\bar{Y}_j) \neq Z_j, \mathrm{m\dot{a}x}|X_{\bar{Y}_j} \cap T_{\pi(\bar{Y}_j)}| \right\}$$

and also we can obtain,

$$\begin{aligned} E_{\pi^*}(N) &= \sum_{v_i \in N} \mathbb{1}\{\pi^*(\bar{Y}_i) \neq Z_i\} \\ &= \sum_{\hat{X}_i \in \hat{X}_{\pi^*}} \sum_{v=(\bar{Y},Z) \in \hat{X}_i} \mathbb{1}\{\pi^*(\bar{Y}) \neq Z\} \\ &= \sum_{\hat{X}_i \in \hat{X}_{\pi^*}} \sum_{v=(\bar{Y},Z) \in \hat{X}_i} \mathbb{1}\{\phi^{-1}(\pi(\bar{Y})) \neq Z\} \end{aligned}$$

Also from Equation 4, we know

$$\mathrm{m\dot{a}x}|X_{\bar{Y}_j} \cap T_{\pi(\bar{Y}_i)}| = \hat{X}_{\pi(\bar{Y}_i)} \cup \underline{X}_{\pi(\bar{Y}_i)}$$

because $T_{\pi(\bar{Y}_i)}$ might already given to bigger intersection so we used $\mathrm{m\dot{a}x}$, and $\underline{X}_{\pi(\bar{Y}_i)}$ define as,

$$\underline{X}_{\pi(\bar{Y}_i)} = \left\{ v_j \in D | \pi(\bar{Y}_i) = \pi(\bar{Y}_j), \pi(\bar{Y}_j) = Z_j, \mathrm{m\dot{a}x}|X_{\bar{Y}_j} \cap T_{\pi(\bar{Y}_j)}| \right\}$$

Based on greedy $\underline{X}_{\pi(\bar{Y}_i)}$ is maximized on other hands from Equation **??**, we know that $E_{\pi^*} \leq E_\pi$, so there exist equivalence C partition based on the ϕ, we have such that using Inequality 2,

$$\sum_{v \in C} \mathbb{1}\{\phi^{-1}(\pi(\bar{Y})) \neq Z\} \leq \sum_{v \in C} \mathbb{1}\{\pi(\bar{Y}) \neq Z\} \tag{4}$$

moreover, this should be true for all $z \in C$. But if $\pi^*(\bar{Y})$ is not $\pi(\bar{Y})$ then,

$$\left| \left\{ v_i \in D ; \mathbb{1}\{\pi^*(\bar{Y}_i) = y_i\} \right\} \right| < \underline{X}_{\pi(\bar{Y}_i)}$$

so this contradicting with Inequality 4 so for equivalence class C, we have

$$\sum_{v \in C} \mathbb{1}\{\pi^*(\bar{Y}) \neq Z\} = \sum_{v \in C} \mathbb{1}\{\pi(\bar{Y}) \neq Z\}$$

and because $\hat{X}_\pi$ and $\hat{X}_{\pi^*}$ is finite, this mean ϕ is identity function $\phi(x) = x$ so $\pi = \pi^*$. That mean greedy algorithm finds the best permutation transformations that satisfies Z. □

2.3 Proof of Lemma 5

Proof. Let $G = (V, E)$, and set Y is the node labels from L assigned to V. Let $C \subseteq L$ be the set of all labels that used in Y. The easy case is when we want to change a color $c \in Y$ to $c' \notin Y$, this is like renaming. To proof this lemma, we use induction. For showing an edge, we use $i + j$ means that two end point of nodes have label i and j and the edge label is $+1$. Let $C = \{c, c'\}$, we have multiple scenarios that generate violation $Vi = \{c' + c, c + c', c - c, c' - c', \}$ and also the set of non-violation scenarios is $nVi = \{c' - c, c - c', c + c, c' + c'\}$ as you can see nVi and Vi closed under swap operation.

We assume the theorem is true for $|C| = k - 1$, let Y used for k colors to color them. We know $k - 1$ colors can swap, only color k is matter now, consider swap $i \in [k - 1]$ and k. All edges involve in this swap is $\{i + k, i - k, k + i, k - i, i + i, i - i, k - k, k + k\}$ and errors involved with these two labels are $\{i + k, k + i, i - i, k - k\}$, and this set size does not change after the swap.

Based on the statement at the beginning of the proof, we are sure about k appear to $[k - 1]$ colors, because it is like renaming, the only thing is changing k to i. Let j be a label such that $e = (v_i, v_j) \in E : \mathit{label}(v_l) = j \bigwedge \mathit{label}(v_m) = k$, the number of error are $\{j + k, k + j\}$ and after swap we have same number of edge in this set. So Y and its version after swap, Y' have same number of edge violations on the label set L, In other word, for any L, we have the following statement. $\sum_{(u,v)\in E} \mathbb{1}\{\varphi(Y_u, Y_v) \neq X_{u,v}\} = \sum_{(u,v)\in E} \mathbb{1}\{\varphi(Y'_u, Y'_v) \neq X_{u,v}\}$ □

2.4 Proof of Lemma 6

Proof. Let $\delta(S)^+$ and $\delta(S)^-$ show the positive and negative edges in $\delta(S)$. We define the external boundary nodes as follow,

$$V^S = \{v \in G : (v, e) \in \delta(S) \wedge v \notin S\}$$

and internal boundary nodes as

$$V_S = \{v \in G : (v, e) \in \delta(S) \wedge v \in S\}$$

It is simple to verify that for each $v \in V^S$ there exist $u \in V_S$ such that $(u, v) \in \delta(S)$ and vice versa. We know that $\tilde{Y}_v^W = Y_v$ for $v \in V^S$. If $\delta(S)^- = \emptyset$ and all edges in $\delta(S)$ be correct, we can follow the labels node in V^S, so for each $v \in V^S$ we select the edges (v, u) in $\delta(S)$ and we define $\mathit{swap}(S, v, u)$ so we have set of mapping $\Phi^+(S) = \{\mathit{swap}(S, v, u) : v \in V^S \wedge u \in V_S \wedge (u, v) \in \delta(S)\}$, from Lemma 5, we know the that the number of violations in S is same, so we resolved some violations in $\delta(S)$ which has contradiction with $\tilde{Y}^W \in \mathbb{I}_{min}$, so when $\delta(S)^- = \emptyset$, at least half of nodes are incorrect and we actually can derive the labeling.

Let $\Gamma_k(S)$ be all label permutation in S such that each permutation can be represented with a sequence of swaps. We can easily show that any sequence of swap is also does not change the edge violation, so we know for all $\pi \in \Gamma_k(S)$ the number of edge violations in S is constant. Because V^S is correct labeled so at least $\lceil \frac{\delta(S)}{2} \rceil$ of edges in $\delta(S)$ are incorrect, otherwise there exist a labeling permutation that contradict with minimization of edge violation because the edges inside S does not add violation but we resolve more than half of $\delta(S)$, In this case we know the existential of such a this permutation but in binary and $\delta(S)^- = \emptyset$ cases, we can actually build the better permutation.

□

2.5 Proof of Lemma 7

Proof. From Lemma 6, we know at least half of $\delta(S)$ for any $S \subset W^*$ are incorrect, so we used this to find an upper bound for this probability, so the best permutation of labels also should satisfy Lemma 6 so we have

$$\mathbb{P}\left(\min_{\pi\in\Gamma_k(W^*)} \mathbb{1}\{\pi(\bar{Y}^{W^*}) \neq Y^W\} > 0\right) \leq \sum_{S\subset W^*, S\cap W\neq\emptyset, \bar{S}\cap W\neq\emptyset} p^{\lceil \frac{\delta(S)}{2} \rceil}$$

$$\leq \sum_{S\subseteq W^*} p^{\lceil \frac{mincut^*(W)}{2} \rceil} \textit{(because } |\delta(S)| \leq mincut^*(W) \textit{ for all } S \subseteq W^*\textit{)}$$

$$\leq 2^{|W^*|} p^{\lceil \frac{mincut^*(W)}{2} \rceil} \quad \textit{(there are } 2^{|W^*|} \textit{ subsets)}$$

where $mincut^*(W) = min_{S \subset W^*, S \cap W \neq \emptyset, \bar{S} \cap W^* \neq \emptyset} |\delta_{G(W)}(S)|$ □

2.6 Proof of Lemma 8

We have following theorem from Boucheron et al. (2003)

Theorem 5. *If there exists a constant $c > 0$ such that $V_+ \leq cS$ then*

$$\mathbb{P}\{S \geq \mathbb{E}[S] + t\} \leq exp\bigg(\frac{-t^2}{4c\mathbb{E}[S] + 2ct}\bigg)$$

Subsequently, with probability at least $1 - \delta$,

$$\begin{aligned} S &\leq \mathbb{E}(S) + \max\left\{4c\log(\frac{1}{\delta}), 2\sqrt{2c\mathbb{E}(S)\log(\frac{1}{\delta})}\right\} \\ &\leq 2\mathbb{E}(S) + 6c\log(\frac{1}{\delta}). \end{aligned}$$

Now we can prove this theorem,

Proof. We define a random variable that shows the number of the component that has an error concerning the real labels of each component. This random variable is a function of given edges X.

$$S(X) = \sum_{W \in \mathcal{W}} \min_{\pi \in \Gamma_k(W)} \mathbb{1}\{\pi(\tilde{Y}^W(X)) \neq Y^W\} \tag{5}$$

We know $S(X) = 0$ means perfect matching with a given X and in maximum $S(X) = |\mathcal{W}|$, and also $\tilde{Y}^W(X)$ is the component-wise estimator with given edge labels observation X. We know that $S : [k]^{|E|} \to \mathbb{R}$ so we can use Theorem 5 if we can prove that $S(X)$ satisfies the assumption.

$$S(X) - S(X^{(e)}) = \sum_{W \in \mathcal{W}} \left(\min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X)) \neq Y^W\Big] - \min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X^{(e)})) \neq Y^W\Big]\right)$$

The right-hand side of the equation is zero for hypernodes that e is not in them so we can reduce the equation to the hypernodes that have e, so we show it with $\mathcal{W}(e)$. Formally $\mathcal{W}(e) = \{W \in \mathcal{W} | e \in E(W)\}$

$$S(X) - S(X^{(e)}) = \sum_{W \in \mathcal{W}(e)} \left(\min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X)) \neq Y^W\Big] - \min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X^{(e)})) \neq Y^W\Big]\right)$$

For evaluate Theorem 5, in next proposition we showed V_+ is bounded.

Proposition. *The variation of V_+ of $S(X)$ in Equation 5 is bounded, $V_+ \leq cS(X)$.*

Proof.

$$\begin{aligned} &(S(X) - S(X^{(e)}))^2 . \mathbb{1}\Big(S(X) > S(X)^{(e)}\Big) = \\ &\mathbb{1}\Big(S(X) > S(X)^{(e)}\Big) \times \sum_{W \in \mathcal{W}(e)} \left(\min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X)) \neq Y^W\Big] - \min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X^{(e)})) \neq Y^W\Big]\right)^2 \\ &\leq \sum_{W \in \mathcal{W}(e)} \left(\min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X)) \neq Y^W\Big]\right)^2 \end{aligned}$$

//second part removed and square of minus part added

$$\leq |\mathcal{W}(e)| \sum_{W \in \mathcal{W}(e)} \min_{\pi \in \Gamma_k(W)} \mathbb{1}\Big[\pi(\tilde{Y}^W(X)) \neq Y^W\Big]$$

Now we can use this for calculating the expectation.

We directly start with V_+ to find its bound.

$$\begin{aligned}
V_+ &= \sum_{e\in E} \mathbb{E}\bigg[(S(X) - S(X^{(e)}))^2 \cdot \mathbb{1}\bigg(S(X) > S(X)^{(e)}\bigg)\bigg| X_1, X_2, \ldots, X_n\bigg] \\
&= \sum_{e\in E} (S(X) - S(X^{(e)}))^2 \cdot \mathbb{1}\bigg(S(X) > S(X)^{(e)}\bigg) \times \mathbb{P}\bigg[(S(X) - S(X^{(e)}))^2 \cdot \mathbb{1}\bigg(S(X) > S(X)^{(e)}\bigg)\bigg| X_1, X_2, \ldots, X_n\bigg]
\end{aligned}$$

(we assume all probabilities are 1)

$$\leq \sum_{e\in E} (S(X) - S(X^{(e)}))^2 . \mathbb{1}\bigg(S(X) > S(X)^{(e)}\bigg)$$

//from last result

$$\begin{aligned}
&\leq \sum_{e\in E} |\mathcal{W}(e)| \sum_{W\in\mathcal{W}(e)} \min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg[\pi(\tilde{Y}^W(X)) \neq Y^W\bigg] \\
&\leq \max_{e\in E} |\mathcal{W}(e)| \sum_{e\in E} \sum_{W\in\mathcal{W}(e)} \min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg[\pi(\tilde{Y}^W(X)) \neq Y^W\bigg] \\
&= \max_{e\in E} |\mathcal{W}(e)| \sum_{W\in\mathcal{W}(e)} \sum_{e\in E} \min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg[\pi(\tilde{Y}^W(X)) \neq Y^W\bigg] \\
&\leq \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)| \sum_{W\in\mathcal{W}(e)} \min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg[\pi(\tilde{Y}^W(X)) \neq Y^W\bigg] \\
&= \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)| S(X)
\end{aligned}$$

Therefore, there is $c = \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)|$ such that $V_+ \leq cS(X)$.

□

So with $c = \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)|$, the Theorem 5 with probability at least $1 - \frac{\delta}{2}$ is valid,

$$S \leq 2\mathbb{E}(S) + 6 \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)| \log(\frac{2}{\delta})$$

We only need to derive $\mathbb{E}(S)$ using Lemma 7, because $\tilde{Y} \in \mathbb{I}_{\tilde{Y}}$, so Lemma 7 is also valid for $\tilde{Y}$,

$$\begin{aligned}
\mathbb{E}(S) &= \sum_{W\in\mathcal{W}} \mathbb{P}\bigg(\min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg\{\pi(\tilde{Y}^W(X)) \neq Y^W\bigg\}\bigg) \times \min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg\{\pi(\tilde{Y}^W(X)) \neq Y^W\bigg\} \\
&= \sum_{W\in\mathcal{W}} \mathbb{P}\bigg(\min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg\{\pi(\tilde{Y}^W(X)) \neq Y^W\bigg\} = 1\bigg) \\
&= \sum_{W\in\mathcal{W}} \mathbb{P}\bigg(\min_{\pi\in\Gamma_k(W)} \mathbb{1}\bigg\{\pi(\tilde{Y}^W(X)) \neq Y^W\bigg\} > 0\bigg) \\
&\leq \sum_{W\in\mathcal{W}} 2^{|W|} p^{\lceil \frac{mincut(W)}{2} \rceil} \text{ // from Lemma 7}
\end{aligned}$$

so finally we have,

$$\min_{\pi\in[\Gamma_k]^{\mathcal{W}}} \sum_{W\in\mathcal{W}} \mathbb{1}\{\pi(\tilde{Y}^W) \neq Y^W\} \leq \sum_{W\in\mathcal{W}} 2^{|W|+1} p^{\lceil \frac{mincut(W)}{2} \rceil} + 6 \max_{e\in E} |\mathcal{W}(e)| \max_{W\in\mathcal{W}} |E(W)| \log(\frac{2}{\delta})$$

□

2.7 Proof of Theorem 2

From Lemma 8, we can directly proof same result for extend of tree components.

Corollary 2. *There is straightforward deduction to derive the result for $W^* = EXT(W)$ on $T = (\mathcal{W}, F)$ with probability $1 - \frac{\delta}{2}$,*

$$\min_{\pi \in [\Gamma_k]^{\mathcal{W}}} \sum_{W \in \mathcal{W}} \mathbb{1}\{\pi(\bar{Y}^{W^*}) \neq Y^{W^*}\} \leq \sum_{W \in \mathcal{W}} 2^{|W^*|+1} p^{\lceil \frac{mincut^*(W)}{2} \rceil} + 6 \max_{e \in E} |\mathcal{W}^*(e)| \max_{W \in \mathcal{W}} |E(W^*)| \log(\frac{2}{\delta})$$

we define the maximum size of a hyper-graph as its degree $deg^*_E(T) = \max_{e \in E} |\mathcal{W}^*(e)|$ which $\mathcal{W}^*(e) = \{W \in \mathcal{W} | e \in E(W^*)\}$ and $E(W^*)$ is the set of all edged in E that are in W^*, so we have

$$\min_{\pi \in [\Gamma_k]^{\mathcal{W}}} \sum_{W \in \mathcal{W}} \mathbb{1}\{\pi(\bar{Y}^{W^*}) \neq Y^{W^*}\} \leq 2^{wid^*(W)+2} \sum_{W \in \mathcal{W}} p^{\lceil \frac{mincut^*(W)}{2} \rceil} + 6 deg^*_E(T) \max_{W \in \mathcal{W}} |E(W^*)| \log(\frac{2}{\delta})$$

Where $wid^*(W) \triangleq \max_{W \in \mathcal{W}} |W^*| - 1$.

Now we can start to Theorem 2,

Proof. To prove this theorem, we need to define a hypothesis class and find information bound for the optimal solution in there, next, we can find a bound for the distance of the real answer of the problem and best answer in the hypothesis class.

Consider the following permutation finding of the components in T:

$$\Pi^\star = \arg \min_{\Pi \in \Gamma_k^{|\mathcal{W}|}} \sum_{W \in \mathcal{W}} \mathbb{1}\{\Pi(\tilde{Y}^W) \neq Y^W\}$$

from Corollary 2, we know that

$$\min_{\pi \in [\Gamma_k]^{\mathcal{W}}} \sum_{W \in \mathcal{W}} \mathbb{1}\{\pi(\tilde{Y}^W) \neq Y^W\} \leq K_n$$

Because $\tilde{Y}^{W^*}$ and $\bar{Y}^{W^*}$ both are in $\mathbb{I}_{\bar{Y}^{W^*}}$ and also $\tilde{Y}^W$ is $\tilde{Y}^{W^*}$ restricted to W and K_n is

$$\begin{aligned} K_n \triangleq & 2^{wid^*(W)+2} \sum_{W \in \mathcal{W}} p^{\lceil \frac{mincut^*(W)}{2} \rceil} + \\ & 6 deg^*_E(T) \max_{W \in \mathcal{W}} |E(W^*)| \log(\frac{2}{\delta}) \end{aligned}$$

So if we have $\Pi^\star$, we can produce a vertex prediction with at most K_n mistakes with probability $1 - \delta$. However, computing $\Pi^\star$ is impossible because we do not have access to Y, so we need to see using Z as a noisy version of Y, how much approximation error will add to the theoretical bound of prediction.

We define the following hypothesis class, which is defined with K_n so we make even bigger to include an even better possible solution.

$$\begin{aligned} \mathcal{F} \triangleq & ([k] \times [k])^{\mathcal{W}} \\ \text{s.t.} & \sum_{(W,W') \in F} \mathbb{1}\{\psi(\pi_W, \pi_{W'}) \neq S(W, W')\} \leq L_n\} \end{aligned}$$

In this context, each element of $([k] \times [k])^{\mathcal{W}}$ is a vector of size $\mathcal{W}$ element which each sown as π. Our goal is to show that best permutation is in $\mathcal{F}$ with high probability.

Such that $L_n = deg(T).K_n$ which enrich the hypothesis class with make it bigger than using K_n. We know that if $\min\limits_{\pi \in [\Gamma_k(W)]} \mathbb{1}\{\pi(\tilde{Y}^W) \neq Y^W\} = 0$ for a component W then we can find a $\bar{\pi}_W \in \Gamma_k$ such that we can effect on Y^W to get $\tilde{Y}^W$ so $\bar{\pi}_W(Y^W) = \tilde{Y}^W$.

We also have

$$\sum_{(W,W')\in F} \mathbb{1}\{\psi(\pi_W, \pi_{W'}) \neq S(W,W')\} = \sum_{(W,W')\in F} \mathbb{1}\{\psi(\pi_W, \pi_{W'}) \neq [2.\mathbb{1}(\tilde{Y}_v^W, \tilde{Y}_v^{W'}) - 1]\}$$

and we know $v \in W \cap W'$, so if for each $W \in \mathcal{W}$ we have $\bar{\pi}_W$, if the range of π_W and $\pi_{W'}$ be same they get 1 and their range is Y, the right hand side also is 1 because the range of two permutations are Y^W and $v \in W \cap W'$, so $\mathbb{1}\{\psi(\pi_W, \pi_{W'}) \neq [2.\mathbb{1}(\tilde{Y}_v^W, \tilde{Y}_v^{W'}) - 1]\} = 0$ when ever W and W' have no errors. Therefore $\Pi^\star \in \mathcal{F}$ with probability $1 - \delta$. The complexity of hypothesis class can parametrized with the size of $\mathcal{F}(X)$ so we have

$$\begin{aligned}
|\mathcal{F}(X)| &= \sum_{m=0}^{L_n} \binom{|\mathcal{W}|}{m} k!^m \\
&\leq \sum_{m=0}^{L_n} \binom{|\mathcal{W}|}{m} k!^{L_n} = k!^{L_n} \sum_{m=0}^{L_n} \binom{|\mathcal{W}|}{m} \\
&\leq k!^{L_n} \left(\frac{e|\mathcal{W}|}{L_n}\right)^{L_n} \leq \left(\frac{e.n.k!}{L_n}\right)^{L_n}
\end{aligned}$$

We consider non-redundant decomposed trees which means for $(W_i, W_j) \in F$ we have $W_i \backslash (W_i \cap W_j) \neq \emptyset$. In Algorithm 3, we use Z instead of Y. So we have

$$\hat{\pi} = \min_{\pi \in \mathcal{F}(X)} \sum_{W \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\big\{\pi(\tilde{Y}_v^W) \neq Z_v\big\}.$$

We have following lemma to continue the proof

Lemma 10. *For* $\sum\limits_{v \in W} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\}$ *we have following approximation,*

$$\begin{aligned}
&\sum_{v \in W} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\} \\
&= \frac{1}{c} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) = Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\}_{=} \\
&+ \frac{1}{c'} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\}_{\neq}
\end{aligned}$$

such that $c = -\big(1 - \frac{k}{k-1}q\big)$ *and* $c' = 1 - \frac{k}{k-1}q$.

Proof. We prove this equation step by step

$$\begin{aligned}
&\sum_{v \in W} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\} = \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) = Y_v} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\}_{=} + \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\}_{\neq} \\
&= \frac{1}{c} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) = Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\}_{=} \\
&+ \frac{1}{c'} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\}_{\neq}
\end{aligned}$$

We have to derive each part of the relation separately, for both sigma if $\hat{\pi}(\tilde{Y}_v^W) = \pi^\star(\tilde{Y}_v^W)$ the above is true for any c and c'.

We need to calculate c and c', for c which is $\pi^\star(\tilde{Y}_v^W) = Y_v$, we have

$$\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) = Y_v\big\} = \frac{k-2}{k-1}q$$

and $\mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v \big| \pi^\star(\tilde{Y}_v^W) = Y_v\big\} = 1-q$ so we can calculate c.

$$\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\} = \frac{k-2}{k-1}q - (1-q) = -\big(1 - \frac{k}{k-1}q\big)$$

so $c = -\big(1 - \frac{k}{k-1}q\big)$. Next, we calculate c' which is $\pi^\star(\tilde{Y}_v^W) \neq Y_v$, therefore we have

$$\begin{aligned}&\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v\big\} =\\ &\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v \wedge Y_v = Z_v\big\} + \mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v \wedge Y_v \neq Z_v\big\} =\\ &\frac{k-2}{k-1}q + 1 - \frac{1}{k-1}q = 1-q\end{aligned}$$

and for second part we have,

$$\begin{aligned}&\mathbb{P}_Z\big\{\pi^\star(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v\big\} =\\ &= \mathbb{P}_Z\big\{\pi^\star(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v \wedge Y_v = Z_v\big\} + \mathbb{P}_Z\big\{\pi^\star(\tilde{Y}_v^W) \neq Z_v \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v \wedge Y_v \neq Z_v\big\} =\\ &= q + \frac{k-2}{k-1}q\end{aligned}$$

so we can calculate c'

$$\begin{aligned}\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\} &=\\ &= (1-q) - \big[q + \frac{k-2}{k-1}q\big]\\ &= 1 - \frac{k}{k-1}q\end{aligned}$$

therefore that $c' = 1 - \frac{k}{k-1}q$.

□

Fix $\hat{\pi} \in \mathcal{F}(X)$ for each component $W \in \mathcal{W}$ we have

$$\begin{aligned}&\sum_{v \in W} \mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W) \neq Y_v\big\} \leq \sum_{v \in W} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\} + \sum_{v \in W} \mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W) \neq Y_v\big\} \qquad \text{//Triangle inequality}\\ &\leq \sum_{v \in W} \mathbb{1}\big\{\hat{\pi}(\tilde{Y}_v^W) \neq \pi^\star(\tilde{Y}_v^W)\big\} + |W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^{W^*}) \neq Y_v\big\} \qquad \text{//Maximize component error}\\ &= -\frac{1}{1 - \frac{k}{k-1}q} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) = Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\}\\ &+ \frac{1}{1 - \frac{k}{k-1}q} \sum_{v \in W \wedge \pi^\star(\tilde{Y}_v^W) \neq Y_v} \Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W) \neq Z_v\big\} - \mathbb{P}_Z\big\{\pi_W^*(\tilde{Y}_v^W) \neq Z_v\big\}\Big\} + |W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W) \neq Y_v\big\} \qquad \text{//From Lemma 10}\end{aligned}$$

For the first part, we can the following approximation:

$$
\begin{aligned}
&-\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)=Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\quad\le 2\sum_{v\in W}\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)=Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\quad\le 2|W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)=Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}
\end{aligned}
$$

We conclude that:

$$
\begin{aligned}
&\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\le 3|W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\\
&\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)\neq Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}+\\
&\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)=Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\le 3|W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\\
&\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)\neq Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}+\\
&\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W\wedge\pi^\star(\tilde{Y}_v^W)=Y_v}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\le 3|W|\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\frac{1}{1-\frac{k}{k-1}q}\sum_{v\in W}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}
\end{aligned}
$$

We apply this formula for all components $W\in\mathcal{W}$ we have

$$
\begin{aligned}
&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\\
&\le 3\Big(\max_{W\in\mathcal{W}}|W|\Big)\sum_{W\in\mathcal{W}}\mathbb{1}\big\{\pi^\star(\tilde{Y}_v^W)\neq Y_v\big\}+\frac{1}{1-\frac{k}{k-1}q}\sum_{W\in\mathcal{W}}\sum_{v\in W}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\le 3\Big(\max_{W\in\mathcal{W}}|W|\Big)K_n+\frac{1}{1-\frac{k}{k-1}q}\sum_{W\in\mathcal{W}}\sum_{v\in W}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}
\end{aligned}
$$

using Lemma 2 for right hand side of the equation, we have excess risk bound with probability $1-\frac{\delta}{2}$,

$$
\begin{aligned}
&\sum_{W\in\mathcal{W}}\sum_{v\in W}\Big\{\mathbb{P}_Z\big\{\hat{\pi}(\tilde{Y}_v^W)\neq Z_v\big\}-\mathbb{P}_Z\big\{\pi^*_W(\tilde{Y}_v^W)\neq Z_v\big\}\Big\}\\
&\le\Big(\frac{2}{3}+\frac{c}{2}\Big)\log(\frac{2|\mathcal{F}(X)|}{\delta})+\frac{1}{c}\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}
\end{aligned}
$$

so we can mix these inequalities,

$$\begin{aligned}&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\\&\leq 3\bigg(\max_{W\in\mathcal{W}}|W|\bigg)K_n+\frac{1}{1-\frac{k}{k-1}q}\bigg(\frac{2}{3}+\frac{c}{2}\bigg)\log(\frac{2|\mathcal{F}(X)|}{\delta})+\frac{1}{c}\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\end{aligned}$$

so we have

$$\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\leq\frac{1}{1-\frac{1}{c}}\Bigg[\bigg(3\max_{W\in\mathcal{W}}|W|\bigg)K_n+\frac{1}{1-\frac{k}{k-1}q}\bigg(\frac{2}{3}+\frac{c}{2}\bigg)\log(\frac{2|\mathcal{F}(X)|}{\delta})\Bigg]$$

We put $c=\frac{1}{1-\epsilon}$ and rearrange then with probability $1-\delta$ we have

$$\begin{aligned}&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\\&\leq\frac{1}{1-\frac{1}{\frac{1}{1-\epsilon}}}\Bigg[\bigg(3\max_{W\in\mathcal{W}}|W|\bigg)K_n+\frac{1}{1-\frac{k}{k-1}q}\bigg(\frac{2}{3}+\frac{\frac{1}{1-\epsilon}}{2}\bigg)\log(\frac{2|\mathcal{F}(X)|}{\delta})\Bigg]\\&=\frac{1}{\epsilon}\Bigg[\bigg(3\max_{W\in\mathcal{W}}|W|\bigg)K_n+\frac{1}{1-\frac{k}{k-1}q}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\log(\frac{2|\mathcal{F}(X)|}{\delta})\Bigg]\end{aligned}$$

From before, we have $|\mathcal{F}(X)|\leq\big(\frac{en.k!}{L_n}\big)^{L_n}$, $wid(T)=\max\limits_{W\in\mathcal{W}}|W|$, K_n, and Lemma 2 so we can conclude

$$\begin{aligned}&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}=\\&=\frac{1}{\epsilon}\bigg(3\max_{W\in\mathcal{W}}|W|\bigg)K_n+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\log(\frac{2|\mathcal{F}(X)|}{\delta})\\&=\frac{3}{\epsilon}.wid(T).K_n+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\times\big(\log(\frac{2}{\delta})+L_n.\log(\frac{en.k!}{L_n})\big))\\&\leq\frac{3}{\epsilon}.wid(T).K_n+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\times\big[\log(\frac{2}{\delta})+K_n.deg(T).k.\log(n.k)\big]\\&=\frac{1}{\epsilon}.K_n\times\big[3.wid(T)+deg(T).k.\log(n.k).\frac{1}{1-\frac{k}{k-1}q}.(\frac{2}{3}+\frac{1}{2(1-\epsilon)})\big]+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\log(\frac{2}{\delta})\\&\leq\frac{1}{\epsilon}.K_n\times\big[3.wid(T)+deg(T).k.\log(n.k).\frac{1}{1-\frac{k}{k-1}q}.(\frac{2}{3}+\frac{1}{2(1-\epsilon)})\big]+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\log(\frac{2}{\delta})\\&\leq\frac{1}{\epsilon}.\Bigg[2^{wid^*(W)+2}\sum_{W\in\mathcal{W}}p^{\lceil\frac{mincut^*(W)}{2}\rceil}+6deg_E^*(T)\max_{W\in\mathcal{W}}|E(W^*)|\log(\frac{2}{\delta})\Bigg]\\&\times\big[3.wid(T)+deg(T).k.\log(n.k).\frac{1}{1-\frac{k}{k-1}q}.(\frac{2}{3}+\frac{1}{2(1-\epsilon)})\big]+\frac{1}{\epsilon.\big(1-\frac{k}{k-1}q\big)}\bigg(\frac{2}{3}+\frac{1}{2(1-\epsilon)}\bigg)\log(\frac{2}{\delta})\end{aligned}$$

so we have

$$\begin{aligned}&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\big\{\hat{\pi}_W(\tilde{Y}_v^W)\neq Y_v\big\}\\&\leq O\Bigg(\frac{1}{\epsilon^2}.\Bigg[2^{wid^*(W)+2}\sum_{W\in\mathcal{W}}p^{\lceil\frac{mincut^*(W)}{2}\rceil}+6deg_E^*(T)\max_{W\in\mathcal{W}}|E(W^*)|\log(\frac{2}{\delta})\Bigg]\times\big[3.wid(T)+deg(T).k.\log(n.k)\big]\Bigg)\end{aligned}$$

because mincut $\geq$ maximum degree

$$\leq\tilde{O}\big(k.\log k.p^{\lceil\frac{\triangleq}{2}\rceil}.n\big)$$

As $\hat{\pi}_W(\tilde{Y}_v) = \hat{Y}_v$, so the algorithm ensures Hamming error has driven upper bound. □

3 MIXTURE OF EDGES AND NODES INFORMATION

In all previous works (Foster et al., 2018; Ofer Meshi & Sontag, 2016; Globerson et al., 2015), the algorithms consider the information of edge and node labels in different stages. For instance in (Globerson et al., 2015), first solves the problem based on the edge because $p < q$, then it uses the nodes information. The information value of positive and negative edges in binary cases are same, but this courtesy breaks under categorical labels, on the other hand, we can use some properties in the graph to trust more on some information. We can calculate the probability of correctness of graph nodes and edges label using p and q. In categorical labeling, the space of noise has some variations from the binary case, so we have the following facts in the categorical case:

- Flipping an edge makes an error.
- Switching the label of a node might not make an error.

Using Bayes rule and the property of nodes, we have $Pr(v = i|v' = j) = Pr(v' = j|v = i))$, the prim for a vertex shows the vertex after effecting noise.

We have following theorem the proof come in supplementary material,

Theorem 6. *The likelihood of correctness of an edge $e = (v_i, v_j) \in E$ with label with L are as follow,*

$$Pr(L \textit{ is untouched } |e, L) =$$
$$c_L \times \begin{cases} 2(1-q)q + (\frac{q}{k-1})^2 . \frac{k-2}{k.(k-1)} & L = 1, vio \\ (1-q)^2 + (\frac{q}{k-1})^2 . \frac{1}{k.(k-1)} & L = 1, nvio \\ 2(1-q). \frac{q}{k-1} + (\frac{q}{k-1})^2 . \frac{k-2}{k(k-1)} & L = -1, vio \\ (1-q)^2 + (\frac{q}{k-1})^2 . \frac{k-2}{k} & L = -1, nvio \end{cases}$$

which $c_L = \frac{(1-p)|E|}{\#L \textit{ in graph}}$, vio means $\phi(X_i, X_j) \neq X_{ij}$, and nvio means $\phi(X_i, X_j) = X_{ij}$.

Proof. In all cases, two head nodes of a given edge are v_i and v_j, and L shows the label of the edge. We first calculate the probability $Pr(v_i, v_j, L|L$ is untouched) the using Bayes theorem, we derive the likelihood.

- The first case is e generates a violation $\phi(Z_i, Z_j) \neq X_{ij}$, and the edge label $L = 1$, in this case, the probability of the event is only one of the node labels are changed or both node labels have been changed but to the different labels.

$$\begin{aligned} &Pr(\text{only one of the node labels are changed}) = \\ &2Pr(v_i \text{ is changed}) = \\ &2(1-q). \sum_{v_i.label=j \wedge j \neq X_i} Pr(v_i.label = j | v_i.label = i) \\ &= 2(1-q). \sum_{v_i.label=j \wedge j \neq X_i} \frac{q}{k-1} = 2.(1-q)q \end{aligned}$$

and also we have, (v'_i and v'_j are the label of given nodes after noise effect)

$$\begin{aligned} &Pr(v'_i \neq v'_j \wedge v_i = v_j \wedge v'_j \neq v_j \wedge v'_i \neq v_i) \\ &= Pr(v_i \neq v'_i).Pr(v_j \neq v'_j).Pr(v_i = v_j) \times Pr(v'_i \neq v'_j | v_i = v_j \wedge v'_j \neq v_j \wedge v'_i \neq v_i) \\ &= \frac{q}{k-1}.\frac{q}{k-1}.\frac{1}{k}.\frac{(k-1)(k-2)}{(k-1).(k-1)} \\ &= (\frac{q}{k-1})^2.\frac{k-2}{k.(k-1)} \end{aligned}$$

Because $Pr(v_i \neq v'_i)$, $Pr(v_j \neq v'_j)$, and $Pr(v_i = v_j)$ are independent, so the whole probability would be $2.(1-q)q + (\frac{q}{k-1})^2.\frac{k-2}{k.(k-1)}$.

- The second case is e does not generate any violation, $\phi(Z_i, Z_j) = X_{ij}$, and the edge label $L = 1$, in this case, either both node labels are untouched or they changed but to the same label.

$$Pr(\text{both node labels are untouched}) =$$
$$Pr(v_i = v'_i).Pr(v_j = v'_j) = (1-q)(1-q) = (1-q)^2$$

  and also we have,

$$\begin{aligned} &Pr(v'_i = v'_j \wedge v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i = v_j) \\ &= Pr(v_i \neq v'_i).Pr(v_j \neq v'_j).Pr(v_i = v_j) \times Pr(v'_i = v'_j | v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i = v_j) \\ &= \frac{q}{k-1}.\frac{q}{k-1}.\frac{1}{k}.\frac{(k-1)(1)}{(k-1).(k-1)} \\ &= (\frac{q}{k-1})^2.\frac{1}{k.(k-1)} \end{aligned}$$

  so the whole probability would be $(1-q)^2 + (\frac{q}{k-1})^2.\frac{1}{k.(k-1)}$.

- The third case is e generates a violation $\phi(Z_i, Z_j) \neq X_{ij}$, and the edge label $L = -1$, in this case, the probability of the event is either one label change to the same label of other head or both change to the same label

$$Pr(\text{a label change to the same of other head})$$
$$= 2Pr(v_i \text{ is changed to } X_j) = 2(1-q).\frac{q}{k-1}$$

  and also we have,

$$\begin{aligned} &Pr(v'_i = v'_j \wedge v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i \neq v_j) \\ &= Pr(v_i \neq v'_i).Pr(v_j \neq v'_j).Pr(v_i \neq v_j) \times Pr(v'_i = v'_j | v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i \neq v_j) \\ &= \frac{q}{k-1}.\frac{q}{k-1}.\frac{k-1}{k}.\frac{(k-2)(1)}{(k-1).(k-1)} \\ &= (\frac{q}{k-1})^2.\frac{k-2}{k(k-1)} \end{aligned}$$

  so the whole probability would be $2(1-q).\frac{q}{k-1} + (\frac{q}{k-1})^2.\frac{k-2}{k(k-1)}$.

- The fourth case is e does not generate any violation, $\phi(Z_i, Z_j) = X_{ij}$, and the edge label $L = -1$, in this case, either both node labels are untouched or they changed but to different labels.

$$\begin{aligned} &Pr(\text{both node labels are untouched}) \\ &= Pr(v_i = v'_i).Pr(v_j = v'_j) \\ &= (1-q)(1-q) = (1-q)^2 \end{aligned}$$

  and also we have,

$$\begin{aligned} &Pr(v'_i \neq v'_j \wedge v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i \neq v_j) \\ &= Pr(v_i \neq v'_i).Pr(v_j \neq v'_j).Pr(v_i \neq v_j) \times Pr(v'_i \neq v'_j | v_i \neq v'_i \wedge v_j \neq v'_j \wedge v_i \neq v_j) \\ &= \frac{q}{k-1}.\frac{q}{k-1}.\frac{k-1}{k}.\frac{(k-1)(k-2)}{(k-1).(k-1)} \\ &= (\frac{q}{k-1})^2.\frac{k-2}{k} \end{aligned}$$

  so the whole probability would be $(1-q)^2 + (\frac{q}{k-1})^2.\frac{k-2}{k}$.

Based on the Bayes theorem we have,

$$Pr(L \text{ is untouched}|v_i, v_j, L) = \frac{Pr(v_i, v_j, L|L \text{ is untouched}).Pr(L \text{ is untouched})}{Pr(v_i, v_j, L)}$$

We have $Pr(v_i, v_j, L) = \frac{\#L \text{ in graph}}{|E|}$, and $Pr(L \text{ is untouched}) = 1 - p$, so we can derive the result. □

As it can be seen with $k = 2$, the trust score for positive and negative are only depend to their frequencies, and if their frequencies are equal we can trust them equally.

Example 1. *(Uniform Frequencies) Let* $\#\{L = +1\} \simeq \#\{L = -1\}$ *and* $k \geq 3$*, then the second part of is negligible because of* $(\frac{q}{k-1})^2$ *parameter, then if* $2(1-q)q \leq (1-q)^2$ *and* $2(1-q).\frac{q}{k-1} \leq (1-q)^2$ *which is* $q < \min\{\frac{1}{3}, \frac{k-1}{k+1}\} = \frac{1}{3}$ *then the non-violating edges are more reliable.*

The following example is more related to the grid graphs that considered in (Globerson et al., 2015).

Example 2. *(Image Segmentation) The case* $k \geq 3$ *and* $\#\{L = +1\} \geq \#\{L = -1\}$*, which we usually see in the images, because the negative edges are on the boundary of regions. If* $q < 1/3$*, We have can trust more on the non-violating negative edges than non-violating positive edges.*

To the best of our knowledge, no algorithm considers the mixture of edges and nodes information on the categorical data. Therefore, Theorem 6 can be a guide to design such an algorithm.

4 EXPERIMENT RESULTS

4.1 Details on Experimental Setup

We provide a detailed discussion on our experimental setup.

Trees Generation Process: We generate random trees, and we apply the noise to the generated graph. We need to have at least one example of each k labels, so the generation process starts by creating k nodes, one example for each category. Then, it generates k random numbers $n_1, \ldots, n_k$ such that $\sum_{i=1}^{k} n_i = n - k$. Next, it creates tree edges for the set of nodes V. Let S and E be empty sets. We select two nodes v and u randomly from V and add (u, v) to E such that the label of the edge satisfies the label of u and v and set $S = S \cup \{u, v\}$, and $V = V \backslash \{u, v\}$. Now, we select one node $v \in S$ and one node $u \in V$ randomly and add (u, v) to E such that the edge label satisfies the endpoints and remove u from V and add it to S. We repeat until V is empty. This process follows the Brooks theorem (Brooks, 1941). Finally, we apply uniform noise model with probabilities of p and q. We select this simple generative process because it covers an extensive range of random trees.

Grids Graph Generation: We use gray scale images as the source of grid graphs. The range of pixel values in gray scale images is $r = [0, 255]$, so we have that $0 \leq k \leq 255$. We divide r to k equal ranges $\{r_1, r_2, \ldots, r_k\}$. We map all pixels whose values are in r_i to $median(r_i)$. For edges, we only consider horizontal and vertical pixels and assign the ground truth edge labels based on the end points. We generate noisy node and edge observations using the uniform noise model. We use Griffin et al. (2007) dataset to select gray-scale images.

Baseline Method: A Majority Vote Algorithm: For each node $v \in G$ assign $f_v = [s_1, s_2, \ldots, s_k]$ with $s_i = 0 : \forall i \in [k]$. Let $label(.)$ shows the label of the passed node. Then, for nodes in neighbourhood of v, $u \in N(v)$, we update f_v with $s_{label(u)} = s_{label(u)} + X_{uv}$. At the end, for each node v, $\hat{Y}_v = \arg\max_{i\in[k]}(f_v)$ if $|\max(f_v)| = 1$ otherwise if $Z_v \in \arg\max(f_v)$, then $\hat{Y}_v = Z_v$ else $\hat{Y}_v = random(\arg\max(f_v))$. This is a simple baseline. We use it as we want to validate that our methods considerably outperform simple baselines.

Evaluation Metric: We use the normalized Hamming distance $\sum_{v\in V} \mathbb{1}(Y_v \neq \hat{Y}_v)/|V|$. between an estimated labeling $\hat{Y}$ and the ground truth labeling Y.

4.2 Additional Experiments on Grids

We provide some qualitative results on the performance of our methods.

Figure 1 presents a qualitative view of the results obtained by our method (and the majority vote baseline) as k increases on the grey scale images. We see that using only the edge information (edge-based prediction) becomes more chaotic for larger values of k. This is because the information that edges carry decreases. However, we see that combining the information provided by both node and edge observations allows us to recover the noisy image. As expected, the simple Majority vote baseline yields worse results than our method.

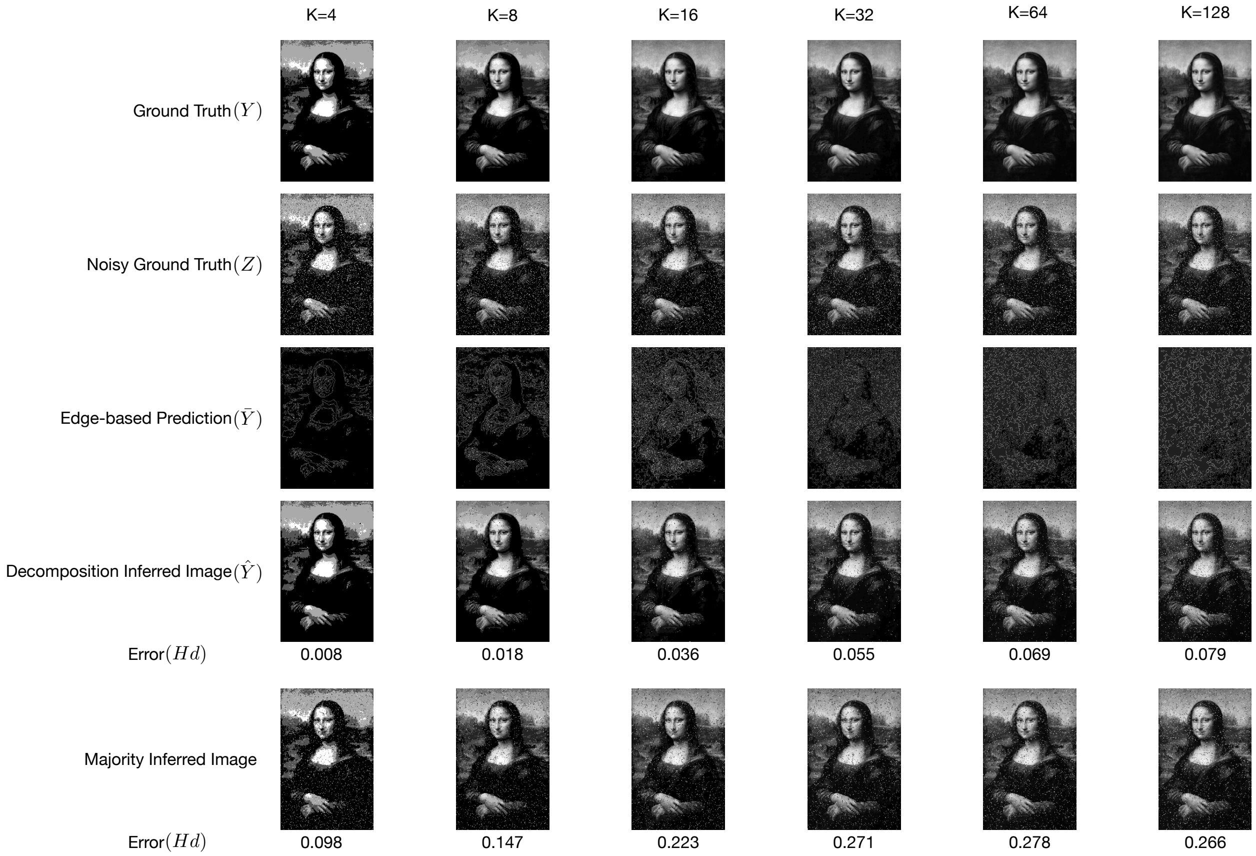

Figure 1: At each column, different stages of the inference process on the image that generates median error can be seen. It starts with generating k value image, adding noise following the model, generates best edge based prediction, and minimize it with noisy ground truth; we also report its corresponding error, you can also see the result and its error from majority algorithm.

5 REFERENCES

Boucheron, S., Lugosi, G., Massart, P., et al. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Brooks, R. L. On colouring the nodes of a network. *Mathematical Proceedings of the Cambridge Philosophical Society*, 37(2):194197, 1941. doi: 10.1017/S030500410002168X.

Foster, D. J., Sridharan, K., and Reichman, D. Inference in sparse graphs with pairwise measurements and side information. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pp. 1810–1818, 2018.

Giotis, I. and Guruswami, V. Correlation clustering with a fixed number of clusters. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1167–1176. Society for Industrial and Applied Mathematics, 2006.

Globerson, A., Roughgarden, T., Sontag, D., and Yildirim, C. How hard is inference for structured prediction? In *International Conference on Machine Learning*, pp. 2181–2190, 2015.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. 2007.

Ofer Meshi, Mehrdad Mahdavi, A. W. and Sontag, D. Train and test tightness of lp relaxations in structured prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1776–1785, 2016.