Approximate Inference in Structured Instances with Noisy Categorical Observations

Alireza Heidari
a5heidar@uwaterloo.ca
Department of Computer Science
University of Waterloo

Ihab F. Ilyas
ilyas@uwaterloo.ca
Department of Computer Science
University of Waterloo

Theodoros Rekatsinas
thodrek@cs.wisc.edu
Department of Computer Science
University of Wisconsin-Madison

Abstract

We study the problem of recovering the latent ground truth labeling of a structured instance with categorical random variables in the presence of noisy observations. We present a new approximate algorithm for graphs with categorical variables that achieves low Hamming error in the presence of noisy vertex and edge observations. Our main result shows a logarithmic dependency of the Hamming error to the number of categories of the random variables. Our approach draws connections to correlation clustering with a fixed number of clusters. Our results generalize the works of Globerson et al. (2015) and Foster et al. (2018), who study the hardness of structured prediction under binary labels, to the case of categorical labels.

1 INTRODUCTION

Statistical inference over structured instances of dependent variables (e.g., labeled sequences, trees, or general graphs) is a fundamental problem in many areas. Examples include computer vision (Nowozin et al., 2011; Dollár & Zitnick, 2013; Chen et al., 2018), natural language processing (Huang et al., 2015; Hu et al., 2016), and computational biology (Li et al., 2007). In many practical setups (Shin et al., 2015; Rekatsinas et al., 2017; Sa et al., 2019; Heidari et al., 2019b), inference problems involve noisy observations of discrete labels assigned to the nodes and edges of a given structured instance and the goal is to infer a labeling of the vertices that achieves low disagreement rate between the correct ground truth labels Y and the predicted labels $\hat{Y}$, i.e., low *Hamming error*. We refer to this problem as *statistical recovery*.

Our motivation to study the problem of statistical recovery stems from our recent work on data cleaning (Rekatsinas et al., 2017; Sa et al., 2019; Heidari et al., 2019b). This work introduces HoloClean, a state-of-the-art inference engine for data curation that casts data cleaning as a structured prediction problem (Sa et al., 2019): Given a dataset as input, it associates each of its cells with a random variable, and uses logical integrity constraints over this dataset (e.g., key constraints or functional dependencies) to introduce dependencies over these random variables. The labels that each random variable can take are determined by the domain of the attribute associated with the corresponding cell. Since we focus on data cleaning, the input dataset corresponds to a noisy version of the latent, clean dataset. Our goal is to recover the latter. Hence, the initial value of each cell corresponds to a noisy observation of our target random variables. HoloClean employs approximate inference methods to solve this structured prediction problem. While its inference procedure comes with no rigorous guarantees, HoloClean achieves state-of-the-art results in practice. Our goal in this paper is to understand this phenomenon.

Recent works have also studied the problem of approximate inference in the presence of noisy vertex and edge observations. However, they are limited to the case of binary labeled variables: Globerson et al. focused on two-dimensional grid graphs and show that a polynomial time algorithm based on MaxCut can achieve optimal Hamming error for planar graphs for which a weak expansion property holds (Globerson et al., 2015). More recently, Foster et al. introduced an approximate inference algorithm based on tree decompositions that achieves low expected Hamming error for general graphs with bounded tree-width (Foster et al., 2018). In this paper, we generalize these results to the case of categorical labels.

Problem and Challenges We study the problem of statistical recovery over categorical data. We consider structured instances where each variable u takes a ground truth label Y_u in the discrete set $\{1, 2, \ldots, k\}$. We assume that for all variables u, we observe a noisy version Z_u of its ground truth labeling such that $Z_u = Y_u$ with probability

$1-q$. We also assume that for all variable pairs (u, v), we observe noisy measurements $X_{u,v}$ of the indicator $M_{u,v} = 2 \cdot \mathbb{1}(Y_u = Y_v) - 1$ such that $X_{u,v} = M_{u,v}$ with probability $1-p$. Given these noisy measurements, our goal is to obtain a labeling $\hat{Y}$ of the variables such that the expected Hamming error between Y and $\hat{Y}$ is minimized. We now provide some intuition on the challenges that categorical variables pose and why current approximate inference methods not applicable:

First, in contrast to the binary case, negative edge measurements do not carry the same amount of information: Consider a simple uniform noise model. In the case of binary labels, observing an edge measurement $X_{u,v} = -1$ and a binary label Z_u allows us to estimate that $\hat{Y}_v = -Z_u$ is correct with probability $(1-q)(1-p) + qp$ when p and q are bounded away from 1/2. However, in the categorical setup, $\hat{Y}_v$ can take any of the $\{1, 2, \ldots, k\} \setminus \{Z_u\}$ labels, hence the probability of estimate $\hat{Y}_v$ being correct is up to a factor of $\frac{1}{k}$ smaller than the binary case. Our main insight is that while the binary case leverages edge labels for inference, approximate inference methods for categorical instances need to rely on the noisy node measurements and the positive edge measurements.

Second, existing approximate inference methods for statistical recovery (Globerson et al., 2015; Foster et al., 2018) rely on a "Flipping Argument" that is limited to binary variables to obtain low Hamming error: for binary node and edge observations, if all nodes in a maximal connected subgraph S are labeled incorrectly with respect to the ground truth, then at least half of the edge observations on the boundary of S are incorrect, or else the inference method would have flipped all node labels in S to obtain a better solution with respect to the total Hamming error. As we discuss later, in the categorical case a naive extension implies that one needs to reason about all possible label permutations over the k labels.

Contributions We present a new approximate inference algorithm for statistical recovery with categorical variables. Our approach is inspired by that of Foster et al. (2018) but generalizes it to categorical variables.

First, we show that, when a variable u is assigned one of the $k-1$ erroneous labels with uniform probability $q/(k-1)$, the optimal Hamming error for trees with n nodes is $\tilde{O}(\log(k) \cdot p \cdot n)$, when $q < 1/2$. This is obtained by solving a linear program using dynamic programming. Here, we derive a tight upper bound on the number of erroneous edge measurements, which we use to restrict the space of solutions explored by the linear program.

Second, we extend our method to general graphs using a tree decomposition of the structured input. We show how to combine our tree-based algorithm with correlation clustering over a fixed number of clusters (Giotis & Guruswami, 2006) to obtain a non-trivial error rate for graphs with bounded treewidth and a specified number of k classes. Our method achieves an expected Hamming error of $\tilde{O}\big(k \cdot \log(k) \cdot p^{\lceil \frac{\Delta(G)}{2} \rceil} \cdot n\big)$ where $\Delta(G)$ is the maximum degree of graph G. We show that local pairwise label swaps are enough to obtain a globally consistent labeling with low expected Hamming error.

Finally, we validate our theoretical bounds via experiments on tree graphs and image data. Our empirical study demonstrates that our approximate inference algorithm achieve low Hamming error in practical scenarios.

2 PRELIMINARIES

We introduce the problem of statistical recovery, and describe concepts, definitions, and notation used in the paper. We consider a structured instance represented by a graph $G = (V, E)$ with $|V| = n$ and $|E| = m$. Each vertex $u \in V$ represents a random variable with ground truth label Y_u in the discrete set $L = \{1, 2, \ldots, k\}$. Edges in E represent dependencies between random variables and each edge $(u, v) \in E$ has a ground truth measurement $M_{u,v} = \varphi(Y_u, Y_v)$ where $\varphi(Y_u, Y_v) = 1$ if $\mathbb{1}(Y_u = Y_v) = 1$ and $\varphi(Y_u, Y_v) = -1$ otherwise.

Uniform Noise Model and Hamming Error We assume access to noisy observations over the nodes and edges of G. For each variable $u \in V$, we are given a noisy label observation Z_u, and for each edge $(u, v) \in E$ we are given a noisy edge observation $X_{u,v}$. These noisy observations are assumed to be generated from G, Y and M by the following process: We are given $G = (V, E)$ and two parameters, edge noise p and node noise $q < 1/2$ with $p < q$. For each edge $(u, v) \in E$, the observation $X_{u,v}$ is independently sampled to be $X_{u,v} = M_{u,v}$ with probability $1-p$ (a *good* edge) and $X_{u,v} = -M_{u,v}$ with probability p (a *bad* edge). For each node $u \in V$, the node observation Z_u is independently sampled to be $Z_u = Y_u$ with probability $1-q$ (a *good* node) and can take any other label in $L \setminus Y_u$ with a uniform probability $\frac{q}{k-1}$. The uniform noise model is a direct extension of that considered by prior work (Globerson et al., 2015; Foster et al., 2018), and a first natural step towards studying statistical recovery for categorical variables.

Given the noisy measurements X and Z over graph $G = (V, E)$, a labeling algorithm is a function A: $\{-1, +1\}^E \times \{1, 2, \ldots, k\}^V \to \{1, 2, \ldots, k\}^V$. We follow the setup of Globerson et al. (2015) to measure the performance of A. We consider the expectation of the Hamming error (i.e., the number of mispredicted labels) over the observation distribution induced by Y. We consider as error the worst-case (over the draw of Y) expected

Hamming error, where the expectation is taken over the process generating the observations X from Y. Our goal is to find an algorithm A such that with high probability it yields bounded worst-case expected Hamming error. In the remainder of the paper, we will refer to the worst-case expected Hamming error as simply Hamming error.

Categorical Labels and Edge Measurements When q is close to 0.5, one needs to leverage the edge measurements to predict the node labels correctly. For binary labels, the structure of the graph G alone determines if one can obtain algorithms with a small error for low constant edge noise p (Globerson et al., 2015; Foster et al., 2018). We argue that this is not the case for categorical labels. Beyond the structure of the graph G, the number of labels k also determines when we can obtain labeling algorithms with non-trivial error bounds.

We use the next example to provide some intuition on how k affects the amount of information in the edge measurements of G: Let nodes take labels in $L = \{1, 2, \ldots, k\}$. We fix a vertex v, and for each vertex u in its neighborhood set the estimate label $\hat{Y}_u$ to Z_u if $M_{u,v} = 1$ and to one of $L \setminus \{Z_u\}$ uniformly at random if $M_{u,v} = -1$. For a correct negative edge measurement and a correct label assignment to v, we are not guaranteed to obtain the correct label for v as we would be able in the binary case.

Given the above setup, the probability that node u is labeled correctly is $P(\hat{Y}_u = Y_u) = (1 - b(1 - \frac{1}{k-1})) \cdot ((1-p)(1-q) + pq))$ where b is the probability of an edge being negative in the ground truth labeling of G. Two observations emerge from this expression: (1) As the number of colors k increases, the probability $P(\hat{Y}_u = Y_u)$ decreases, hence, for a fixed graph G as k increases, statistical recovery becomes harder; (2) For a fixed graph G, as k increases the probability b of obtaining a negative edge in the ground truth labeling of G increases— this holds for a fixed graph G and under the assumption that each label should appear at least once in the ground truth— and the term $(1 - b(1 - \frac{1}{k-1}))$ approaches zero. This implies that for $P(\hat{Y}_u = Y_u)$ to be meaningful the term $((1-p)(1-q) + pq)$ should be maximized for fixed q, and hence, the edge noise p should approach zero as a function of $(1 - b(1 - \frac{1}{k-1}))$. In other words, p should be upper bounded by a function $\phi(k)$ such that as k increases $\phi(k)$ goes to zero. We leverage these two observations to specify when statistical recovery is possible.

Statistical Recovery Statistical recovery is possible for the family $\mathcal{G}$ of structured instances with k categories, if there exists a function $f(p, k) : [0, 1] \to [0, 1]$ with $\lim_{p\to 0} f(p, k) = 0$ such that for every p that is upper bounded by a function $\phi(k)$ with $\lim_{k\to V} \phi(k) = 0$, the Hamming error of a labeling algorithm on graph $G \in \mathcal{G}$ with $V = n$ vertices is at most $f(p, k) \cdot n$.

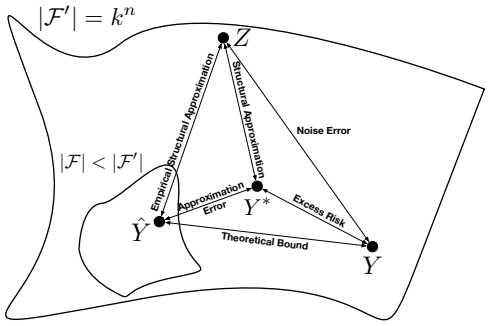

Figure 1: A schematic overview of our approach. Given the noise node labeling Z of a graph with ground truth labeling Y, we leverage the noisy side information to obtain an approximate labeling $\hat{Y}$. Labeling $\hat{Y}$ is an approximate solution to the information theoretic optimal solution Y^*. The goal of our analysis is to find a theoretical bound on the Hamming error between $\hat{Y}$ and Y.

3 APPROACH OVERVIEW

We consider a graph $G = (V, E)$ with node labels in $L = \{1, 2, \ldots, k\}$. The space of all possible labelings of V defines a *hypothesis space* $\mathcal{F}'$. In this space, we denote Y the latent, ground truth labeling of G. In the absence of any information the size of this space is $|\mathcal{F}'| = k^n$. Access to any side information allows us to identify a subspace of $\mathcal{F}'$ that is close to Y.

First, we consider access only to noisy node labels of G and denote Z the point in $\mathcal{F}'$ for this labeling. If we have no side information on the edges of G, the information theoretic optimal solution to statistical recovery is Z (because we assume $q < 1/2$). Second, we assume access only to edge measurements for G. We denote X the observed edge measurements. If the edge measurements are accurate (i.e., $p = 0$) the size of $\mathcal{F}'$ reduces to $k!$. We assume that k is such that one can obtain a labeling for G that is edge-compatible with X by traversing G. Under this assumption, the number of edge-compatible labelings is equal to all possible label permutations, i.e., $|\mathcal{F}'| = k!$. Finally, in the presence of both node and edge observations the information theoretic optimal solution to statistical recovery corresponds to a point Y^* that is obtained by running exact marginal inference (Globerson et al., 2015). However, exact inference can be intractable, and even when it is efficient, it is not clear what is the optimal Hamming error that Y^* yields with respect to Y.

To address these issues, we propose an approximate inference scheme and obtain a bound on the worst-case expected Hamming error that it obtains. We start with the noisy edge observations X and use them to find a subspace $\mathcal{F} \subset \mathcal{F}'$ that contains node labelings which induce edge labelings that are close to X (in terms of Hamming distance). We formalize this in the next two sections. In-

tuitively, we have that noisy edge measurements partition the space $\mathcal{F}$ in a collection of *edge classes*.

Definition 1. *The edge class of a point $Y \in \mathcal{F}$ is a set $\mathbb{I} \in 2^{\{1,2,\ldots,k\}^{|V|}}$ such that for all $Y_i \in \mathbb{I}$, Y_i induces the same edge measurements as Y. All points in $\mathbb{I}$ can be derived via a label permutation of Y. In general, for any labeling Y', set $\mathbb{I}_{Y'}$ is the set of all labelings that can be generated by a label permutation of Y'.*

The restricted subspace $\mathcal{F}$ contains those edge classes that are close to the noisy edge observations X.

Given the restricted subspace $\mathcal{F}$, we design an algorithm to find a point $\hat{Y} \in \mathcal{F}$ such that the Hamming error between $\hat{Y}$ and Y^* is minimized. We define the Hamming error with respect to an edge class $\mathbb{I}$ as:

Definition 2. *The Hamming error of a vector $\mathcal{Q} \in \{1,2,\ldots,k\}^{|V|}$ to the edge class $\mathbb{I}_{Y'} \in 2^{\{1,2,\ldots,k\}^{|V|}}$ is $Hd(\mathcal{Q}, \mathbb{I}_{Y'}) = \min_{\mathcal{Y} \in \mathbb{I}_{Y'}} Hd(\mathcal{Y}, \mathcal{Q})$.*

Point Y^* might not be in $\mathcal{F}$ and the distance between $\hat{Y}$ and Y^* is the approximation error we have due to approximate inference. Finally, we prove that the expected Hamming error between $\hat{Y}$ and Z is bounded. A schematic diagram of our approximate inference method is shown in Figure 1. In the following sections, we study statistical recovery for trees (in Section 4) and general graphs (in Section 5). All proofs can be found in the supplementary material of our paper (Heidari et al., 2019a).

4 RECOVERY IN TREES

We focus on trees and introduce a linear program for statistical recovery over k-categorical random variables. We prove that under a uniform noise model the optimal Hamming error is $\tilde{O}(\log(k) \cdot p \cdot n)$.

4.1 A Linear Program for Statistical Recovery

We follow the steps described in Section 3. First, we use the noisy edge observations to restrict the search for $\hat{Y}$ to a subspace $\mathcal{F}$. We describe $\mathcal{F}$ via a constraint on the number of edge disagreements between the edge labeling implied by $\hat{Y}$ and the noisy edge observations X. Second, we form an optimization problem to find a point $\hat{Y}$ with minimum Hamming distance from Z that satisfies the aforementioned constraint.

The ground truth edge labeling M (corresponding to the ground truth node labeling Y) has bounded Hamming distance from the observed noisy labeling X. Hence, we can restrict the space of considered solutions to node labelings that induce an edge labeling with a bounded Hamming distance from the observed noisy labeling X. We have: Under the uniform noise model, edge measurements are flipped independently. Thus, the total number of bad edges is a sum over independent and identically distributed (iid) random variables. The expected number of flipped edges is $p \cdot |E| = p(n-1)$. Using the Bernstein inequality, we have:

Lemma 1. *Let G be a graph with noisy edge observations with noise parameter p. With probability at least $1-\delta$ over the draw of X:*

$$\sum_{(u,v)\in E} \mathbb{1}\{\varphi(Y_u, Y_v) \neq X_{u,v}\} \leq t \quad \textit{where}$$

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

This lemma states that under the uniform noise model the ground truth edge labeling M for Graph G is in the neighborhood of X with high probability. Given this bound, we use the following linear program to find $\hat{Y}$:

$$\begin{aligned} \min_{\hat{Y} \in [k]^{|V|}} \quad & \sum_{v \in V} \mathbb{1}\{\hat{Y}_v \neq Z_v\} \\ \text{s.t.} \quad & \\ & \sum_{(u,v)\in E} \mathbb{1}\{\varphi(\hat{Y}_u, \hat{Y}_v) \neq X_{u,v}\} \leq t \end{aligned} \tag{1}$$

where t is defined as in Lemma 1. This problem can be solved via a dynamic programming algorithm with cost $O(k \cdot n^3 \cdot p)$. We describe this algorithm in the supplementary material of the paper (Heidari et al., 2019a).

Discussion Our approach is similar to that of Foster et al. (2018) for binary random variables. However, we use the Bernstein inequality to obtain a tighter concentration bound on the number of flipped edge measurements. In the case of categorical random variables, it is critical to obtain a tight description of the space $\mathcal{F}$ of the possible labeling solutions as we have a larger hypothesis space.

Let $\mathcal{S}(n,k)$ be the size of hypothesis space with k labels and n nodes. If we increase n by one, the rate of change for the hypothesis space is $r_{k,n} = \Delta\mathcal{S}/\Delta n = k^n(k-1)$, which is multiplicative with respect to k. Similarly, as we increase k to $k+1$ the size of the hypothesis space changes by $s_{k,n} = \Delta\mathcal{S}/\Delta k = \sum_{i+j=n-1}(k+1)^i k^j \geq k^{n-1}$, which is exponential in the size of our input. We need a tight bound to obtain an efficient dynamic programming algorithm with respect to n and k.

4.2 Upper Bound on the Hamming Error for Trees

The Hamming error of $\hat{Y}$ obtained by Linear Program 1 is bounded by $\tilde{O}(\log(k) \cdot p \cdot n)$ with high probability. For our analysis, we draw connections to statistical learning.

We define a hypothesis class $\mathcal{F}$ that contains all points that satisfy the bound in Lemma 1:

$$\mathcal{F} = \{Y' \in [k]^{|V|} : \sum_{(u,v)\in E} \mathbb{1}\{\varphi(Y'_u, Y'_v) \neq X_{u,v}\} \leq t\}$$

From Lemma 1, we have that the edge class that corresponds to the ground truth labeling Y is contained in $\mathcal{F}$ with high probability over the draw of X. Moreover, since the node noise q is bounded away from $1/2$, we can use the noisy node measurements Z to find a labeling $\hat{Y}$ that is in the same edge class as Y and close to Y. Such a labeling is obtained by solving Linear Program 1. From a statistical learning perspective, $\hat{Y}$ corresponds to the *empirical risk minimizer* (ERM) over $\mathcal{F}$ given Z. Thus, the Hamming error between $\hat{Y}$ and Y is associated with the *excess risk* over Z for Class $\mathcal{F}$. We have:

Lemma 2. *(Foster et al., 2018) Let $\hat{Y}$ be the empirical risk minimizer over $\mathcal{F}$ given Z and let $Y^* = \arg\min_{Y'\in\mathcal{F}} \sum_{v\in V} \mathbb{P}(Y'_v \neq Y_v)$ and $c > 0$ a constant number, then with probability $1-\delta$ over the draw of Z,*

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

We now analyze how the Hamming error relates to excess risk for categorical random variables. We have:

Lemma 3. *The Hamming error is proportional to the excess risk: For fixed $\hat{Y}, Y \sim \mathcal{F}'$ and Z distributed according to the uniform noise model we have that:*

$$\mathbb{1}\{\hat{Y}_v \neq Y_v\} = \frac{1}{c}\Big[P_Z(\hat{Y}_v \neq Z_v) - P_Z(Y_v \neq Z_v)\Big] \text{ where}$$
$$c = 1 - k/(k-1)q$$

With $k = 2$ we have that $c = 1 - 2q$, which recovers the result of Foster et al. (2018) for binary random variables.

Using Lemma 2, we can bound the excess risk in terms of the size of the hypothesis class. We have:

Corollary 1. *When $Y \in \mathcal{F}$ and $\hat{Y} = \arg\min_{\mathcal{Y}\in\mathcal{F}} \sum_{v\in V} \mathbb{1}\{\mathcal{Y}_v \neq Z_v\}$, we have that with probability at least $1-\delta$ over the draw of Z:*

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

We now combine these results with the complexity of class $\mathcal{F}$ to obtain a bound for the Hamming error:

Theorem 1. *Let $\hat{Y}$ be the solution to Problem 1. Then with probability at least $1-\delta$ over the draw of X and Z*

$$\sum_{v\in V} \mathbb{1}\{\hat{Y}_v \neq Y_v\} \leq \frac{[t\log(2k) - \log(\delta)]}{\big(1 - \frac{k}{k-1}q\big)} \left(\frac{4}{3} + \frac{2}{\frac{1}{4} + \big(\frac{1}{4} - \epsilon\big)\big(1 - \frac{k}{k-1}\big)}\right) = \tilde{O}(\log(k)np)$$

Here, t is the same as in Lemma 1. We see that k has a lower impact on the Hamming error than n and p. Also, when $k = 2$ we recover the result of Foster et al. (2018). Due to the tools we use to prove this result, this is a tight bound. We validate this bound empirically in Section 6.

5 RECOVERY IN GENERAL GRAPHS

We now show how our tree-based algorithm can be combined with correlation clustering to obtain a non-trivial error rate for graphs with bounded treewidth and k-categorical random variables. We first describe our approximate inference algorithm and then show that our algorithm achieves an expected Hamming error of $\tilde{O}\big(k \cdot \log(k) \cdot p^{\lceil \frac{\Delta(G)}{2} \rceil} \cdot n\big)$ where $\Delta(G)$ is the maximum degree of the structured instance G.

5.1 Approximate Statistical Recovery

We build upon the concept of *tree decompositions* (Diestel, 2018). Let G be a graph, T be a tree, and $\mathcal{W} = (V_t)_{t\in T}$ be a family of vertex sets $V_t \subseteq V(G)$ indexed by the nodes t of T. We denote a tree-decomposition with $(T, \mathcal{W})$. The width of $(T, \mathcal{W})$ is defined as $\max\{|V_t| - 1 : t \in T\}$ and the treewidth $tw(G)$ of G is the minimum width among all possible decompositions of G. We also denote with F the $|\mathcal{W}| - 1$ edges connecting the bags in $\mathcal{W}$ in $(T, \mathcal{W})$ and represent T as $T = (\mathcal{W}, F)$.

Given a graph G, a tree decomposition of T defines a series of local subproblems whose solutions can be combined via dynamic programming to obtain a global solution for the original problem on G. For graphs of bounded treewidth, this approach allows us to obtain efficient algorithms (Bodlaender, 1988). Our solution proceeds as follows: Let $(T, \mathcal{W})$ be a tree decomposition of G. We first find a *local* labeling $\tilde{Y}^W$ for each $W \in \mathcal{W}$. Then, we design a dynamic programming algorithm that combines all local labelings to obtain a global labeling $\hat{Y}$.

5.1.1 Finding Local Labelings

We recover the labeling of the nodes in a bag W as follows: (1) Given W, we consider a superset of W, defined

as $W^* = EXT(W) = W \cup \left(\bigcup_{v \in G} N(v)\right)$ where $N(v)$ is the one-hop neighborhood of node v; (2) Given W^*, we use the edge observations in the edge subset $E' \subseteq E$ induced by W^* to find a restricted hypothesis space $\mathcal{F}_{W^*}$. We then find a labeling $\tilde{Y}^{W^*} \in \mathcal{F}_{W^*}$ that has the minimum Hamming error with respect to Z for the nodes in W^*. Let Z_{W^*} denote this subset of Z; (3) For W, we assign $\tilde{Y}^W$ to be the restriction of $\tilde{Y}^{W^*}$ on W.

We consider two cases for Step 2 from above: (1) If $|W^*| = O(\log(n))$, we can enumerate all $k^{O(\log(n))}$ labelings for W^* and choose the one with minimum Hamming distance from Z. The complexity of this brute-force algorithm is $k^{O(\log(n))} = \texttt{poly}(n)$; (2) If $|W^*| = \Omega(\log(n))$, we use the MAXAGREE$[k]$ algorithm of Giotis & Guruswami (2006) over the noisy edge measurements X to restring the subspace $\mathcal{F}$ in the neighborhood of X. MAXAGREE$[k]$ is a polynomial-time approximation scheme (PTAS) for solving the Max-Agreement version of correlation clustering for a fixed number of k labels. In the worst case, MAXAGREE$[k]$ obtains an approximation of 0.7666OPT$[k]$. In our analysis, we account for the approximation factor 0.7666 by changing the probability p to $p' = 0.7666p + 0.2334$. A detailed discussion is provided in the supplementary material of the paper (Heidari et al., 2019a). Given the output of MAXAGREE$[k]$, let $\mathcal{F}_{CC}$ be the restricted subspace of solutions for W^*. We pick an arbitrary labeling $\bar{Y}^{W^*} \in \mathcal{F}_{CC}$ and use Algorithm 1 to get a permutation that transforms $\bar{Y}^{W^*}$ to point $\tilde{Y}^{W^*}$ that has minimum Hamming distance to Z^{W^*}.

Algorithm 1 Local Label Permutation

Input: A labeling $\bar{Y}^{W^*}$ in the subspace $\mathcal{F}_{CC}$ identified by MAXAGREE$[k]$ on W^*; Node observations Z^{W^*};
$\bar{Y}_1^{W^*}, \bar{Y}_2^{W^*}, \dots \bar{Y}_k^{W^*} \leftarrow$ Group $\bar{Y}^{W^*}$ By Label;
$Z_1^{W^*}, Z_2^{W^*}, \dots Z_k^{W^*} \leftarrow$ Group Z^{W^*} By Label;
for $i, j \in [k] \times [k]$ **do**
  $I_{i,j} \leftarrow |\bar{Y}_i^{W^*} \cup Z_j^{W^*}|$;
end for
$Q \leftarrow$ A queue that sorts $I = \{I_{i,j}\}_{(i,j) \in [k] \times [k]}$ in decreasing order with respect to values $I_{i,j}$;
while $Q \neq \emptyset$ **do**
  $I_{i,j} \leftarrow$ Pop(Q);
  $\pi(i) \leftarrow j$;
  Remove all $I_{t,j}$ and $I_{i,t}$ for all $t \in [k]$ from Q;
end while
Return: π

Algorithm 1 greedily permutes the labels in $\bar{Y}^w$ to obtain a labeling with minimum Hamming distance to Z^W. The complexity of this algorithm is $O(n + k \log k)$.

Lemma 4. *Algorithm 1 finds a permutation π such that:*

$$\tilde{Y}^W = \pi(\bar{Y}^W) = \min_{\pi \in \Gamma_k} \sum_{v \in W} \mathbb{1}\{\pi(\bar{Y}^W) \neq Z^W\}$$

where Γ_k is the set of all permutations of the k labels.

We combine all steps in Algorithm 2. The output of this algorithm is a collection of labelings $\tilde{Y}$ for the local problems. Lemma 4 states that $\tilde{Y}^{W^*}$ minimizes the Hamming distance to Z. We also show that $\tilde{Y}^{W^*}$ remains a minimizer with respect to $\min_y \sum_{(u,v)} \mathbb{1}(\varphi(y_u, y_v) \neq X_{uv})$ after the swaps due to π.

Algorithm 2 Find Local Labelings

Input: A tree decomposition $T = (\mathcal{W}, F)$ of G; Noisy node observations Z; Noisy edge measurements X;
$\tilde{Y} \rightarrow \emptyset$;
for $W \in \mathcal{W}$ **do**
  $W^* = EXT(W)$;
  $\backslash *$ The next optimization problem can be solved either via enumeration or correlation clustering. $E(W^*)$ denotes the set of edges in $W^*.* \backslash$
  $\bar{Y}^{W^*} = \arg\min_y \sum_{(u,v) \in E(W^*)} \mathbb{1}\{\varphi(y_u, y_v) \neq X_{uv}\}$;
  $\tilde{Y}^{W^*} \leftarrow$ Local Label Permutation $(\bar{Y}^{W^*}, Z^{W^*})$;
  Let $\tilde{Y}^W$ be the restriction of $\tilde{Y}^{W^*}$ to W;
  $\tilde{Y} \rightarrow \tilde{Y} \cup \{\tilde{Y}^W\}$;
end for
Return: $\tilde{Y}$

Definition 3. *Given a graph $G = (V, E)$, the $swap(V, c_1, c_2)$ function changes all node labels c_1 to c_2, and all node labels c_2 to c_1.*

The swap operation enables us to switch between elements within an edge class. We show that a $swap(V, c_1, c_2)$ does not affect the disagreements between the node labeling and edge labeling of a graph.

Lemma 5. *Let L be a set of labels $L = \{1, 2, \dots, k\}$. Consider a graph $G = (V, E)$ for which we are given a node labeling Y and an edge labeling X. For any pair $(c, c') \in L \times L$, let $Y' = swap(V, c, c')$ be the node labeling of G after swapping label c with c'. We have that: $\sum_{(u,v) \in E} \mathbb{1}\{\varphi(Y_u, Y_v) \neq X_{u,v}\} = \sum_{(u,v) \in E} \mathbb{1}\{\varphi(Y'_u, Y'_v) \neq X_{u,v}\}$.*

This lemma implies that $\tilde{Y}^{W^*}$ is a minimizer of $\min_y \sum_{(u,v)} \mathbb{1}\{\varphi(y_u, y_v) \neq X_{uv}\}$ since $\bar{Y}^{W^*}$ minimizes this quantity, and $\tilde{Y}^{W^*}$ is a permutation of $\bar{Y}^{W^*}$.

5.1.2 From Local Labelings to a Global Labeling

We now describe how to combine labelings $\{\tilde{Y}^W\}_{W \in \mathcal{W}}$ into a global labeling $\hat{Y}$. For binary random variables, the following procedure plays a central role in enforcing agreement across local labelings (Foster et al., 2018): Given a bag W_1 and a neighbor W_2 with conflicting node labels with respect to W_1, we can maximize the agreement between W_1 and W_2 by flipping labeling $\tilde{Y}^{W_1}$ to its mirror labeling. This operation leads to consistent

solutions since for binary random variables there is only one mirror labeling. However, for categorical random variables we have $k!$ possible mirror labelings for $\tilde{Y}^{W_1}$. *We show that it suffices to consider only one label swap per bag instead of $k!$ labelings.*

We consider the swap operation (see Section 5.1.1) and two bags W_1 and W_2 with labelings $\tilde{Y}^{W_1}$ and $\tilde{Y}^{W_2}$. We resolve conflicts in $W_1 \cap W_2$ as follows: Let $\Pi_k \subset \Gamma_k$ be the set of all permutations restricted to one pairwise color swap. Given a bag $W \in \mathcal{W}$ with labeling Y^W, we define a swap $\pi = swap(W, c_i, c_j)$ to be valid if color c_i is present in Y_W. Given a valid swap π for W, we define $\pi(Y^W)$ to be the label assignment for all nodes in W after applying π to Y^W. Also, let $\pi(Y_v^W)$ be the labeling for a node $v \in W$ after π. Finally, we define $\Pi_k(Y^W)$ as the set of all labelings for W that can be obtained if we apply any valid pairwise label swap on Y^W. To resolve inconsistencies between $\tilde{Y}^{W_1}$ and $\tilde{Y}^{W_2}$, we consider pairs in $\Pi_k(Y^{W_1}) \times \Pi_k(Y^{W_2})$ such that the labeling in the intersection of W_1 and W_2 is consistent and the number of nodes whose label is swapped is minimum.

The procedure we use is shown in Algorithm 3. The algorithm takes as input a tree decomposition $T = (\mathcal{W}, F)$ of G and the local labelings $\tilde{Y}$. For each W with labeling $\tilde{Y}^W$, we compute the cost of swapping label c_i with label c_j for each $(i, j) \in [k] \times [k]$. Then, we iterate over edges in F to identify incompatibilities between local node labelings. Finally, we use all the computed costs to find the single swap π_W to be applied locally to each bag $W \in \mathcal{W}$ such that global agreement is maximized. To this end, we solve a linear program similar to program 1. This program is shown in Algorithm 4.

In Algorithm 4, function $\psi(\cdot)$ is defined as:

$$\psi(\pi_W, \pi_{W'}) = \\ = \begin{cases} 1, \text{ if } \pi_W(\tilde{Y}_v^W) = \pi_{W'}(\tilde{Y}_v^{W'}) : \forall v \in W \cap W' \\ -1, \text{ if } \pi_W(\tilde{Y}_v^W) \neq \pi_{W'}(\tilde{Y}_v^{W'}) : \exists v \in W \cap W' \end{cases}$$

Constant L_n is used to restrict the space of solutions considered. A discussion on L_n is deferred to Section 5.2.

5.1.3 Discussion on Correlation Clustering

We use correlation clustering in our algorithm for practical reasons. If the cardinality of the bags $T = (\mathcal{W}, F)$ is bounded by $O(\log(n))$, we can find a local labeling for each W that has minimum Hamming distance to Z efficiently. Obtaining such a decomposition T is an NP-complete problem. This challenge is also highlighted by Foster et al. (2018). To address this issue they assume a sampling procedure for removing edges from G to obtain a subgraph for which a low-width tree decomposition is easy to find. This procedure is a graph-specific exercise and not easily generalizable to arbitrary graphs. We follow a different approach. Instead of using specialized procedures, we rely on heuristics to obtain a low-width decompositions de Givry et al. (2006); Dermaku et al. (2008) and use correlation clustering for large bags. This scheme allows us to use our algorithm with arbitrary graphs.

Algorithm 3 From Local Labelings to a Global Labeling

Input: A tree decomposition $T = (\mathcal{W}, F)$ of G; Noisy node observations Z; Noisy edge measurements X; Local labelings $\{\tilde{Y}^W\}_{W \in \mathcal{W}}$;
$\hat{Y} \to \emptyset$;
for $W \in \mathcal{W}$ **do**
  $\Pi_k^W \leftarrow$ the set of valid pairwise color swaps for W;
  for $\pi \in \Pi_k^W$ **do**
    $\setminus * \pi$ is associated with a label swap (c_i, c_j) $* \setminus$;
    $Cost_W[\pi] = \sum_{v \in W} \mathbb{1}(\pi(\tilde{Y}^W) \neq Z^W)$;
  end for
end for
for $(W_1, W_2) \in F$ **do**
  Select one node v from $W_1 \cap W_2$ randomly;
  $S(W_1, W_2) = 2 \cdot \mathbb{1}\{\tilde{Y}_v^{W_1} = \tilde{Y}_v^{W_2}\} - 1$;
end for
Compute constant L_n; $\setminus *$ See Section 5.2 $* \setminus$;
$\{\pi_W\}_{W \in \mathcal{W}} =$ Cat. Tree Decoder$(T, Cost, S, L_n)$;
for $v \in V$ **do**
  Choose arbitrary W s.t. $v \in W$ randomly;
  $\hat{Y}_v = \pi_W(\tilde{Y}_v^W)$
end for
Return: $\hat{Y}$

Algorithm 4 Categorical Tree Decoder

Input: A tree $T = (\mathcal{W}, F)$; Matrices $\{Cost_W\}_{W \in \mathcal{W}}$, $\{S(W, W')\}_{(W,W') \in F}$, $L_n \in \mathbb{N}$;
Output: Optimal swaps $\{\pi_W\}_{W \in \mathcal{W}}$ for each $W \in \mathcal{W}$;
Solve the linear program:

$$\hat{\Pi} = \arg \min_{\{\pi_W\}_{W \in \mathcal{W}} \in \Pi_k^{|\mathcal{W}|}} \sum_{W \in \mathcal{W}} Cost_W[\pi_W]$$

$$\text{s.t.} \sum_{(W,W') \in F} \mathbb{1}\{\psi(\pi_W, \pi_{W'}\} \neq S(W, W') \leq L_n$$

Return: $\hat{\Pi}$

5.2 A Bound for Low Treewidth Graphs

We state our main theorem for statistical recovery over general graphs. We also provide a proof sketch.

Theorem 2. (**Main Theorem**) *Consider graph G with $T = (\mathcal{W}, F)$, noisy node observations Y, and noisy edge observations X. Let $\hat{Y}$ be the statistical recovery solution obtained by combining Algorithms 2 and 3. With high probability over the draw of Z and X:*

$$\sum_{v \in V} \mathbb{1}\{\hat{Y}_v \neq Y_v\} \leq \tilde{O}\big(k \cdot \log k \cdot p^{\lceil \frac{\text{mincut}^*(G)}{2} \rceil} \cdot n\big) \\ \leq \tilde{O}\big(k \cdot \log k \cdot p^{\lceil \frac{\Delta}{2} \rceil} \cdot n\big)$$

where $mincut^*(G)$ is the min. mincut over all extended bags in $\mathcal{W}$ and $\Delta(G)$ is the max. degree in G.

We see that the Hamming error obtained by our approach goes to zero as $p \to 0$. Theorem 2 allows us to understand when statistical recovery over a graph with categorical random variables is possible (i.e., when we can rely on edge observations to solve statistical recovery more accurately than the trivial solution of keeping the initially assigned node labels). Theorem 2 connects the level of edge-noise with the degree Δ of the input graph, the number of labels k, and the noise q on node labels. We have that for the edge noise p it should be $p \leq \sqrt[\lceil \frac{\Delta}{2} \rceil]{\frac{q}{k \log k}}$, where q is the node noise parameter, for the side information in X to be useful for statistical recovery. Otherwise, one should just use the initially observed node labels.

Proof Sketch Let S denote a maximal connected subgraph of G. Let $\delta(S)$ be the boundary of S, i.e., the set of edges with exactly one endpoint in S. Let $\tilde{Y}^S$ be the local labeling for nodes in S. We say that S is incorrectly labeled if for all $v \in S$ we have $\tilde{Y}_v^S \neq Y_v$. We have:

Lemma 6. *(Swapping lemma) Let S be a maximal connected subgraph of G with every node incorrectly labelled by $\tilde{Y}$. Then at least half the edges of $\delta(S)$ are bad.*

For a bag W, let set S be the largest connected component in W such that for all nodes v in it $\tilde{Y}_v^W \neq Y_v$. It must be the case that at least half of the $\delta(S)$ edges are incorrect or else there exists a different labeling that agrees with X better than $\tilde{Y}^W$. This contradicts the fact that $\tilde{Y}^W$ is a minimizer of $\min_y \sum_{(u,v)} \mathbb{1}\{\varphi(y_u, y_v) \neq X_{uv}\}$. This result extends the Flipping Lemma of Globerson et al. (2015) from the binary to the categorical case.

We use this result to bound the probability that a local labeling $\tilde{Y}^W$ (see Lemma 4) will fail to recover the ground truth node label for W. The probability of local labelings having large Hamming error is upper bounded:

Lemma 7. *Let Γ_k be the all label permutations on the set $L = \{1, 2, \ldots, k\}$. We have for W:*

$$\mathbb{P}\left(\min_{\pi \in \Gamma_k} \mathbb{1}\{\pi(\tilde{Y}^W) \neq Y^W\} > 0 \right) \leq 2^{|W^*|} p^{\lceil \frac{mincut^*(W)}{2} \rceil}$$

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

We now build upon Lemma 7 and leverage the result introduced by Boucheron et al. (2003) to obtain an upper bound on the total number of mislabeled nodes across all bags in $\mathcal{W}$ for any labeling permutation $\pi \in \Gamma_k$ over the local labeling $\tilde{Y}^W$:

Lemma 8. *Let Γ_k be the all label permutations on the set $L = \{1, 2, \ldots, k\}$. For all $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$ over the draw of X we have that:*

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

$$2^{|W|+1} p^{\lceil \frac{\text{mincut}(W)}{2} \rceil} + 6 \max_{e \in E} |\mathcal{W}(e)| \max_{W \in \mathcal{W}} |E(W)| \log(\frac{2}{\delta})$$

where $\mathcal{W}(e)$ denotes the set of bags in $\mathcal{W}$ that contain edge e and $E(W)$ denotes the set of edges in bag W.

This lemma can be extended to W^* as well. This lemma combined with Lemma 6 implies that the labeling disagreement across bags in the tree decomposition are bounded. The analysis continues in a way similar to that for trees (see Section 4). Given the local bag labelings, we seek to find the labeling swaps across bags such that the global labeling has minimum Hamming error with respect to Y. We use the inequality from Lemma 8 to restrict the space $([k] \times [k])^{\mathcal{W}}$ of all possible pairwise label swaps over the local bag labelings. Let s^* be the optimal point in $([k] \times [k])^{\mathcal{W}}$ such that the global labeling has minimum Hamming error with respect to Y. Given the tree decomposition $T = (\mathcal{W}, F)$ of G. We define the hypothesis space:

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

with $L_n = deg(T)\big[2^{wid^*(W)+2} \sum_{W \in \mathcal{W}} p^{\lceil \frac{mincut^*(W)}{2} \rceil} + 6 deg_E^*(T) \max_{W \in \mathcal{W}} |E(W^*)| \log(\frac{2}{\delta})\big]$, $deg_E^*(T) = \max_{e \in E} |\mathcal{W}(e)|$, and π_W and $S(W, W')$ denote the pairwise swaps and labeling disagreements between bags from Algorithm 3. We show that the optimal permutation Π^* is a member of $\mathcal{F}$ with high probability and also have that $|\mathcal{F}(X)| \leq \big(\frac{e \cdot n \cdot k!}{L_n}\big)^{L_n}$. Combining this with Lemma 2, we take $\hat{\Pi}$ is most correlated with Z, i.e., it is a minimizer for $\sum_{W \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\big\{\pi_W(\tilde{Y}_v^W) \neq Z_v\big\}$. Directly from statistical learning theory we have that the Hamming error of this estimator $\hat{Y}$ is $\tilde{O}(\log(\mathcal{F})) = \tilde{O}\big(k \cdot \log k \cdot p^{\lceil \frac{\Delta}{2} \rceil} \cdot n\big)$ which establishes our main theorem.

6 EXPERIMENTS

Experimental Setup We evaluate our approach on trees and grid graphs. For trees, we use Erdős–Rényi random trees to obtain ground truth instances. For grids, we use real images to obtain the ground truth. We create noisy observations via a uniform noise model. We compare our approach with two approximate inference baselines: (1) a Majority Vote algorithm, where we leverage the neighborhood of a node to predict its label, and (2) (Loopy) Belief Propagation. To evaluate performance we use the

normalized Hamming distance $\sum_{v \in V} \mathbb{1}(Y_v \neq \hat{Y}_v)/|V|$. We provide more details in the Supplementary Material.

Hamming Error of Random Trees Our analysis suggests that Linear Program 1 yields a solution with Hamming error $\tilde{O}(\log(k)np)$. We evaluate experimentally that the Hamming error increases at a logarithmic rate with respect to k. Figure 2 shows the Hamming error for a fixed tree generative model with $p = 0.1$ and $q = 0.2$ as we increase the number of labels k. We fix q away from 0.5 and generate $10,000$ trees for each k. We report the average error. As shown, we observe the expected logarithmic behavior that we proved theoretically. The graph size is chosen randomly $n \in [10^3, 1.5 \times 10^3]$.

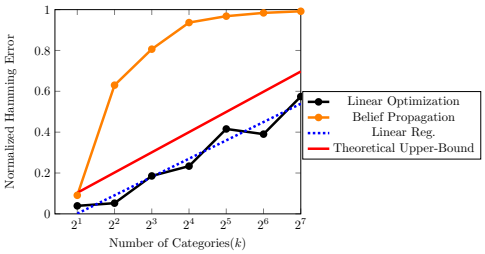

Figure 2: Experimental validation that Hamming error for trees increases with a logarithmic rate w.r.t. k.

Hamming Error of Grids We have two experiments on grids. In the first experiment, we select $1,000$ grayscale images and compute the Hamming error obtained by our algorithm. We consider a uniform noise model with $p = 0.05$ and $q = 0.1$. Figure 3 shows the Hamming error as k increases. As expected we see that the Hamming error increases. This is because as k increases negative edges carry lower information, and with non-zero edge error (p), the positive edges also provide low information observations (i.e., a wrong measurement). In the supplementary material of our paper, we present a qualitative evaluation of our results on the grey-scale images.

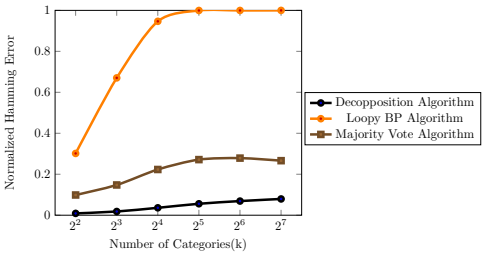

Figure 3: The Hamming error for different methods on grids. We show mean the mean error of $1,000$ repetitions.

In the second experiment, we evaluate the effect of edge noise p on the quality of solution obtained by our methods for a fixed number of labels k and fixed node noise q. In Figure 4, we show the effect of p on the average of Hamming error when other parameters are fixed ($n = 6 \times 10^4, k = 128, q = 0.1$). We vary p from zero to 0.5. We repeat each experiment 100 times. We find that our approximate inference algorithm is robust to small amounts of noise.

This experiment also validates Theorem 2 which states when the side information from edges X helps with statistical recovery. For the setups we consider in this experiment, we have k = 128 and vary q in 0.1, 0.15, 0.2. If we keep the initial node labels the expected normalized Hamming error will be 0.1, 0.15, and 0.2 respectively. Theorem 2 states that to obtain a better Hamming error than the above one, the edge noise p has to be less than $\sqrt{0.1/(128 \log 128)} \sim 0.04$, $\sqrt{0.15/(128 \log 128)} \sim 0.05$, $\sqrt{0.2/(128 \log 128)} \sim 0.06$ respectively. Figure 4 shows that the normalized Hamming error obtained by our algorithm reaches the Hamming error of the trivial algorithm (and plateaus around it) at the expected edge-noise levels of 0.04, 0.05, and 0.06.

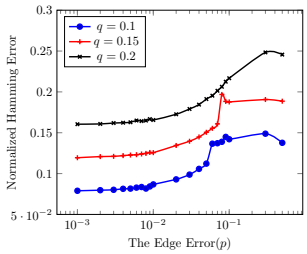

Figure 4: The effect of varying p on the average of normalized Hamming error(Hd) with fixed q.

7 CONCLUSION

We considered the problem of statistical recovery in structured instances with noisy categorical observations. We presented a new approximate algorithm for inference over graphs with categorical random variables. We showed a logarithmic dependency of the Hamming error to the number of categories the random variables can obtain. We also explored the connections between approximate inference and correlation clustering with a fixed number of clusters. There are several future directions suggested by this work. One interesting direction would be to understand under which noise models the problem of statistical recovery is solvable. Moreover, it is interesting to explore the direction of correlation clustering further and extend our analysis beyond small tree width graphs.

Acknowledgements The authors thank Shai Ben David, Fereshte Heidari Khazaei, and Joshua McGrath for the helpful discussions. This work was supported by Amazon under an ARA Award, by NSERC under a Discovery Grant, and by NSF under grant IIS-1755676.

References

Bodlaender, H. L. Dynamic programming on graphs with bounded treewidth. In Lepistö, T. and Salomaa, A. (eds.), *Automata, Languages and Programming*, pp. 105–118. Springer Berlin Heidelberg, 1988.

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

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

de Givry, S., Schiex, T., and Verfaillie, G. Exploiting tree decomposition and soft local consistency in weighted csp. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pp. 22–27. AAAI Press, 2006. ISBN 978-1-57735-281-5.

Dermaku, A., Ganzow, T., Gottlob, G., Mcmahan, B., Musliu, N., and Samer, M. Heuristic methods for hypertree decomposition. In *Proceedings of the 7th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*, MICAI '08, pp. 1–11, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88635-8.

Diestel, R. *Graph theory*. Springer Publishing Company, Incorporated, 2018.

Dollár, P. and Zitnick, C. L. Structured forests for fast edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1841–1848, 2013.

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.

Heidari, A., Ilyas, I. F., and Rekatsinas, T. Approximate inference in structured instances with noisy categorical observations. *arXiv preprint arXiv:2749231*, 2019a.

Heidari, A., McGrath, J., Ilyas, I. F., and Rekatsinas, T. Holodetect: Few-shot learning for error detection. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, pp. 829–846, 2019b.

Hu, Z., Ma, X., Liu, Z., Hovy, E., and Xing, E. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318*, 2016.

Huang, Z., Xu, W., and Yu, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Li, M.-H., Lin, L., Wang, X.-L., and Liu, T. Protein–protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5):597–604, 2007.

Nowozin, S., Lampert, C. H., et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4): 185–365, 2011.

Rekatsinas, T., Chu, X., Ilyas, I. F., and Ré, C. Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11):1190–1201, 2017.

Sa, C. D., Ilyas, I. F., Kimelfeld, B., Ré, C., and Rekatsinas, T. A formal framework for probabilistic unclean databases. ICDT, 2019.

Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., and Ré, C. Incremental knowledge base construction using deepdive. *Proceedings of the VLDB Endowment*, 8(11): 1310–1321, 2015.