A APPENDIX

A.1 Proof of Theorem 3.1

We begin by introducing some notation from prior work [1]. First, recall that a degree sequence for a relation R and variable V is a vector $f_{R.V}$ such that $f_{R.V,i}$ is the frequency of the ith most frequent value, and $F_{R.V,i} = \sum_{j=1,i} f_{R.V,j}$ is the cumulative degree sequence (CDS). We represent relations R as tensors M_R with one dimension for each variable present in the relation. The value at a particular entry $M_{i_1,\ldots,i_{|V_R|}}$ is equal to the frequency of that tuple, $(i_1,\ldots,i_{|V_R|})$, in the relation R. Note that entries can be zero if that tuple does not appear in R. Additionally, the discrete derivative and integral of a tensor M on a variable V is defined as,

Definition A.1.

$$(\Delta_V M)_{v_i} = M_{v_i} - M_{v_i-1} \tag{1}$$

$$(\Sigma_V M)_{v_i} = \sum_{j=1,v_i} M_j \tag{2}$$

A tensor M is *consistent* with a set of degree sequences f_R if the following is true,

$$(\sum_{V' \neq V} M)_i \leq f_{R.V,i} \qquad\qquad \forall\, i \in \mathbb{D}_V, V \in V_R$$

Briefly, this means that if we contract the tensor down to a single dimension, then the resulting vector is less than the DF $F_{R.V}$ at all points. The set $\mathcal{M}_{f_R}$ is the set of tensors consistent with f_R.

Lastly, we define the *value tensor*, E^{F_R}, and use it to explicitly define the *worst-case tensor*, C^{F_R}.

Definition A.2. The *value* tensor, $E^{F_R} \in \mathbb{R}_+^{[n]}$, is defined by the following linear optimization problem:

$$\forall m \in [n]: \qquad\qquad E_m^{F_R} \stackrel{\text{def}}{=} \text{Maximize:} \sum_{s \leq m} M_s \tag{3}$$

$$\text{Where: } M \in \mathcal{M}_{f_R}$$

The *worst-case* tensor, $C^{F_R} \in \mathbb{R}^{[n]}$, is defined as:

$$C^{f_R} \stackrel{\text{def}}{=} \Delta_{V_1} \cdots \Delta_{V_d} E^{F_R} \tag{4}$$

Or, equivalently,

$$\Sigma_{V_1} \cdots \Sigma_{V_d} C^{f_R} = E^{F_R} \tag{5}$$

Note that this worst-case tensor is equivalent to the worst-case instance, $W(R)$ of a relation R, as depicted in Figure 2.

Getting back to Theorem 3.1 of this work, we start by proving that it holds for star queries before expanding to all Berge-acyclic queries. Consider the following query,

$$Q_{STAR} = R(V_1,\ldots,V_d)S_1(V_1)\ldots S_d(V_d)$$

If M is the count tensor of the relation R and $a^{(V_i)}$ is the count tensor of S_i, in this case a simple non-increasing vector, then we can express the query size as follows,

$$|Q_{STAR}(D)| = M \cdot a^{(V_1)} \cdots a^{(V_d)}$$

Given this notation, we consider part of Theorem 3.2 from [1].

Theorem A.3. *[Thm. 3.2 from [1]] Let f_R be the set of degree sequences as above, and let V, C defined by (3)-(5). Then:*

    (1) We can define the value tensor as follows,

$$\forall m \in [n]: \qquad\qquad E_m^{F_R} = \min\left(F_{R.V_1}(m_1),\ldots,F_{R.V_d}(m_d)\right) \tag{6}$$

    (2) For any non-increasing vectors $a^{(V_p)} \in \mathbb{R}_+^{[n_p]}$, $p = 2, d$, the vector $C^{f_R} \cdot a^{(V_2)} \cdots a^{(V_d)}$ is in $\mathbb{R}_+^{[n_1]}$ and non-increasing.

    (3) For all count tensors M_R, and all non-increasing vectors $a^{(X_1)} \in \mathbb{R}_+^{[n_1]},\ldots,a^{(X_d)} \in \mathbb{R}_+^{[n_d]}$:

$$M_R \cdot a^{(V_1)} \cdots a^{(V_d)} \leq C^{f_R} \cdot a^{(V_1)} \cdots a^{(V_d)} \tag{7}$$

    Directly implying,

$$|Q_{STAR}(D)| \leq |Q_{STAR}(W(s(D)))| \tag{8}$$

Let $\hat{F}_R$ be an upper bound of F_R, i.e. $\hat{F}_{R.V}(i) \geq F_{R.V}(i) \,\forall\, V \in V_R, i$, and define $\hat{f}_{R.V} = \Delta_V \hat{F}_{R.V}$ and $\hat{f}_R$ as the set of these degree sequences which, as specified in Theorem 3.1, must be non-increasing. Further, note that item 1 and item 1 relies only on the properties of the worst-case instance's inherent structure, so it immediately applies to $C^{\hat{F}_R}$.

Based on the above, we can prove the following lemma,

Lemma A.4. *For all non-increasing vectors $\boldsymbol{a}^{(V_1)} \in \mathbb{R}_+^{[n_1]}, \ldots, \boldsymbol{a}^{(V_d)} \in \mathbb{R}_+^{[n_d]}$:*

$$C^{f_R} \cdot \boldsymbol{a}^{(V_1)} \cdots \boldsymbol{a}^{(V_d)} \leq C^{\hat{F}_R} \cdot \boldsymbol{a}^{(V_1)} \cdots \boldsymbol{a}^{(V_d)} \tag{9}$$

Directly implying,

$$|Q_{STAR}(W(s(D)))| \leq |Q_{STAR}(W(\Delta\hat{S}))| \tag{10}$$

PROOF. Following the original proof of item 2, we begin by simplifying the problem using 1-0 vectors. In particular, let $\boldsymbol{b}^{(m)} \in \mathcal{R}^n$ be the vector with m 1's followed by $n-m$ 0's. Because the $\boldsymbol{a}^{V_i}$ are non-increasing integral vectors, they can be represented as a sum of 1-0 vectors, so it suffices to consider the case where each of them is a 1-0 vector. In this case, the problem description becomes,

$$C^{f_R} \cdot \boldsymbol{b}^{(m_1)} \cdots \boldsymbol{b}^{(m_d)} \leq C^{\hat{F}_R} \cdot \boldsymbol{b}^{(m_1)} \cdots \boldsymbol{b}^{(m_d)}$$

Multiplying against $\boldsymbol{b}^{(m)}$ is the same as summing over the first m indices, so this can be alternatively expressed as,

$$\Sigma_{m_1} \ldots \Sigma_{m_d} C^{f_R} \leq \Sigma_{m_1} \ldots \Sigma_{m_d} C^{\hat{f}_R}$$

Considering the definition of the value tensor $E_{\boldsymbol{m}}^{f_R}$, we can rephrase this as follows where $\boldsymbol{m} = (m_1, \ldots, m_d)$,

$$E_{\boldsymbol{m}}^{f_R} \leq E_{\boldsymbol{m}}^{\hat{f}_R}$$

Lastly, we insert the alternative definition of E^{f_R} provided in item 1 and the fact that each $\hat{F}_{R.V_i}$ is an upper bound of $F_{R.V_i}$ to prove the lemma,

$$\min(F_{R.V_1}(m_1), \ldots, F_{R.V_d}(m_d)) \leq \min(\hat{F}_{R.V_1}(m_1), \ldots, \hat{F}_{R.V_d}(m_d))$$

□

To prove that this can be extended to general queries, we rely on more theory from [1].

Theorem A.5. *Implied by Thm. 4.2 of [1] If the following is true for a set of database instances, $\mathcal{D}$,*

(1) *For any non-increasing vectors $\boldsymbol{a}^{(V_p)} \in \mathbb{R}_+^{[n_p]}, p = 2, d$, the vector $C^{f_R} \cdot \boldsymbol{a}^{(V_2)} \cdots \boldsymbol{a}^{(V_d)}$ is in $\mathbb{R}_+^{[n_1]}$ and non-increasing.*

(2) *For all relations $R \in \mathcal{D}$ with count tensor $\boldsymbol{M}_R$, and all non-increasing vectors $\boldsymbol{a}^{(X_1)} \in \mathbb{R}_+^{[n_1]}, \ldots, \boldsymbol{a}^{(X_d)} \in \mathbb{R}_+^{[n_d]}$:*

$$\boldsymbol{M}_R \cdot \boldsymbol{a}^{(V_1)} \cdots \boldsymbol{a}^{(V_d)} \leq C^{f_R} \cdot \boldsymbol{a}^{(V_1)} \cdots \boldsymbol{a}^{(V_d)} \tag{11}$$
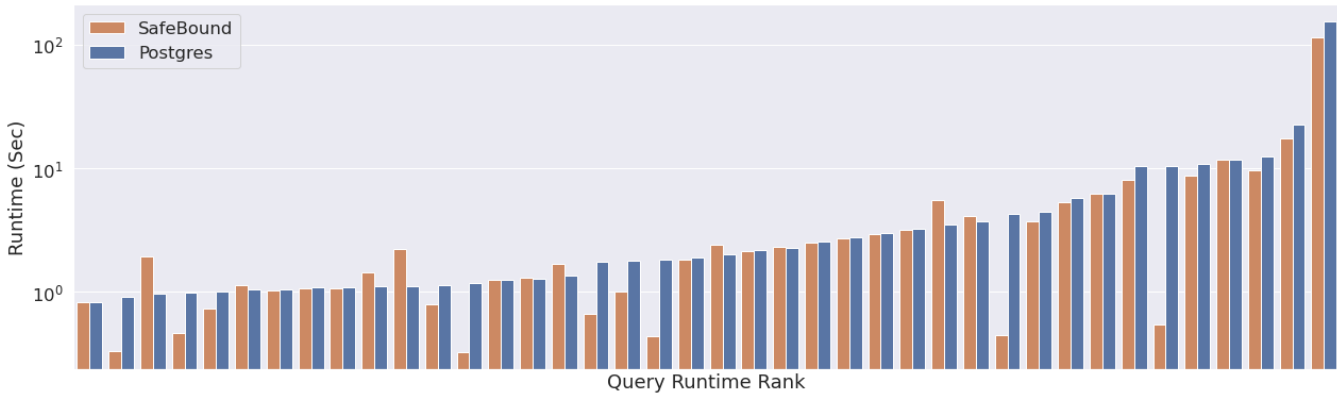
Then, for any Berge-acyclic query, Q,

$$|Q(D)| \leq |Q(W(s))| \quad \forall D \in \mathcal{D} \tag{12}$$

The first immediately holds for $C^{\hat{F}_R}$ from Theorem A.3 while the latter holds for the set of all database instances, D, such that $D \models S$ due to Lemma A.4.
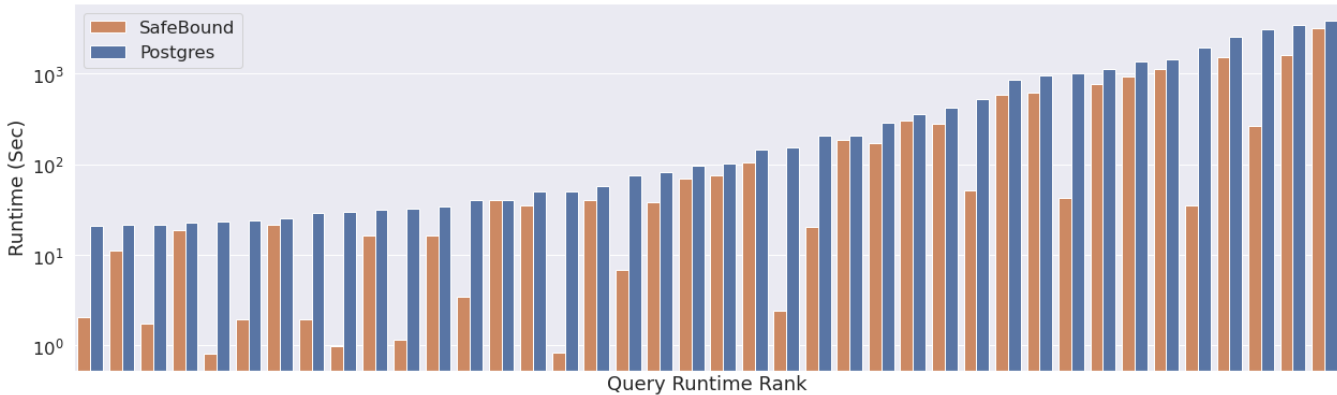
B   DISAGGREGATED RESULTS

In this section, we provide alternate versions of Figures 6, 7, and 9.a which are disaggregated by benchmark.
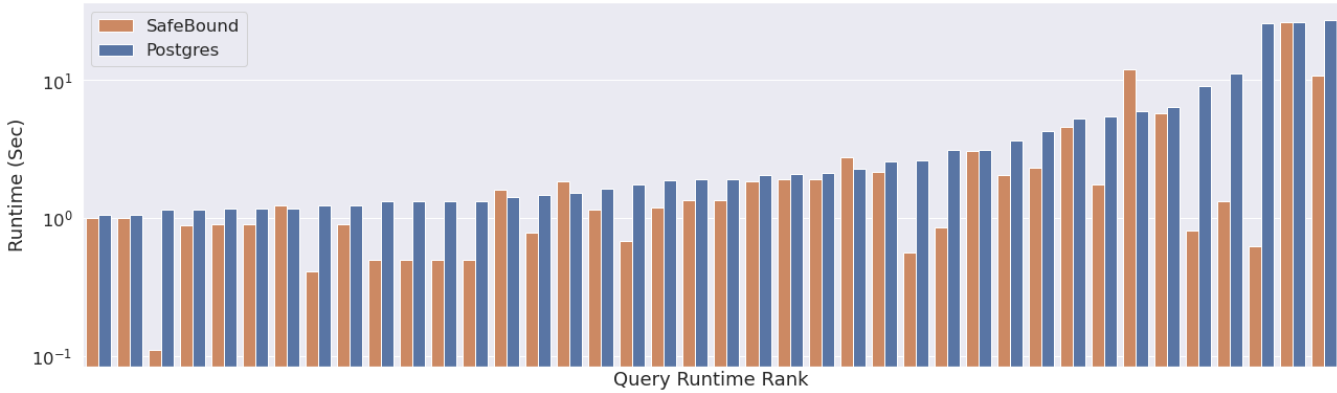
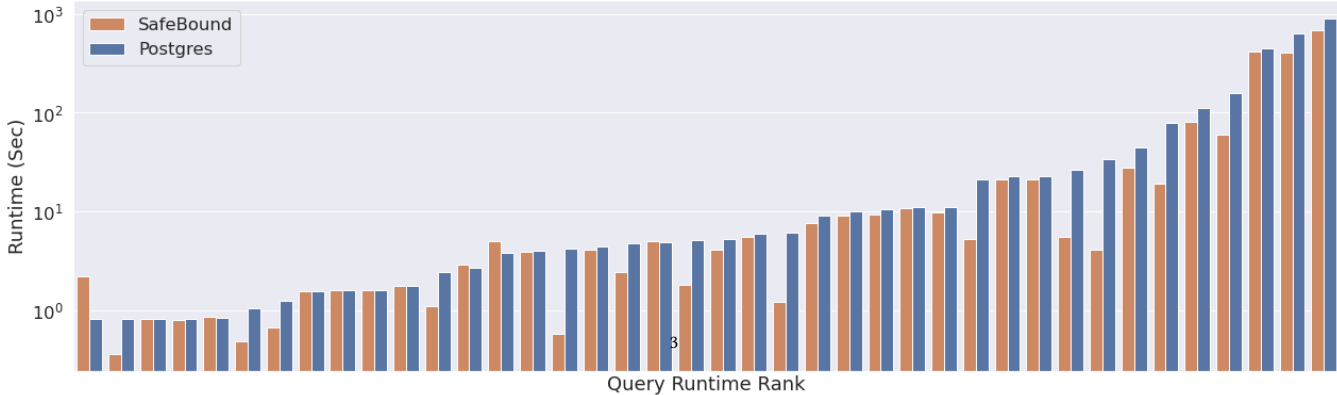## C   RELATIVE ERROR BY JOIN SIZE

(a) Top 40 Query Runtimes - JOBLight

(b) Top 40 Query Runtimes - JOBLightRanges

(c) Top 40 Query Runtimes - JOBM

(d) Top 40 Query Runtimes - Stats

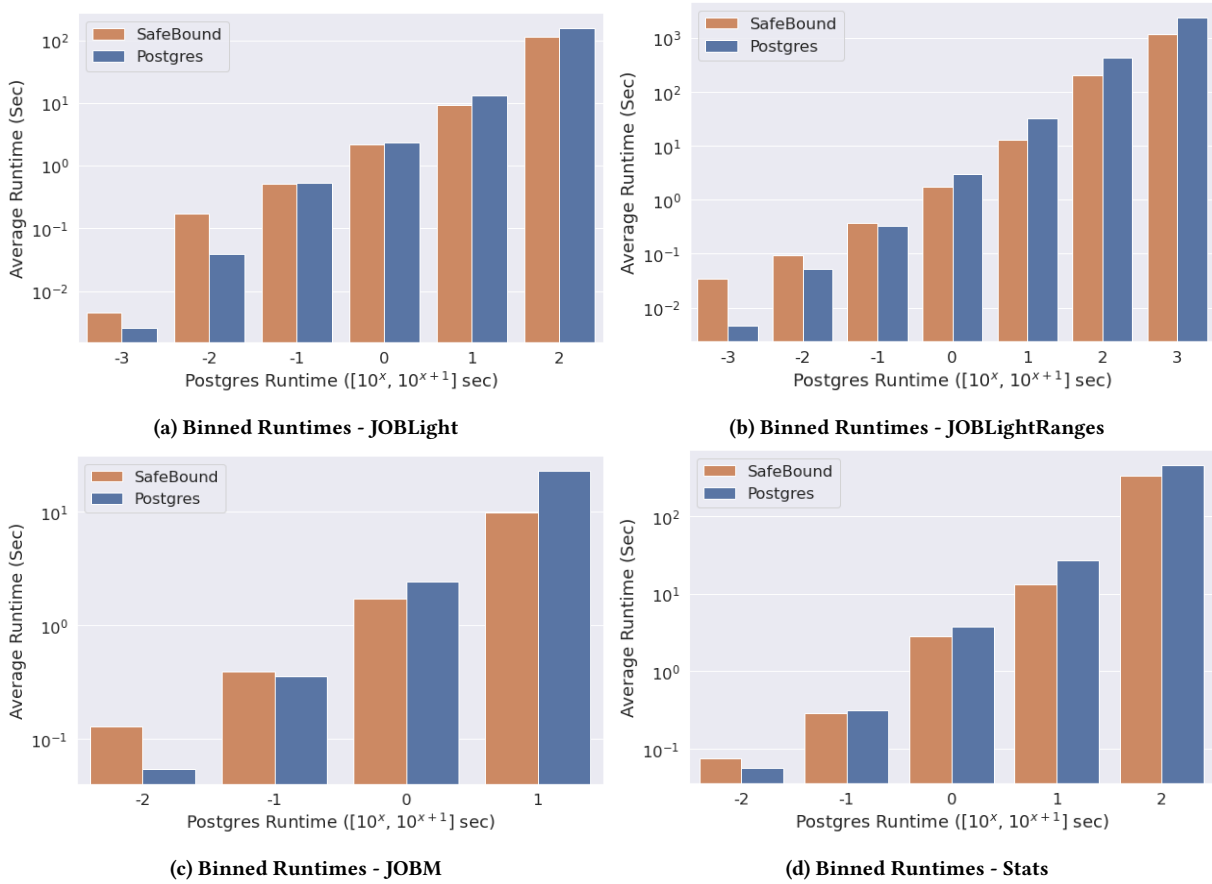Figure 1: Top Query Runtime Graphs

(a) Binned Runtimes - JOBLight

(b) Binned Runtimes - JOBLightRanges

(c) Binned Runtimes - JOBM

(d) Binned Runtimes - Stats

Figure 2: Binned Runtime Graphs

REFERENCES

[1] Kyle Deeds, Dan Suciu, Magda Balazinska, and Walter Cai. 2022. Degree Sequence Bound For Join Cardinality Estimation. *CoRR* abs/2201.04166 (2022). arXiv:2201.04166 https://arxiv.org/abs/2201.04166
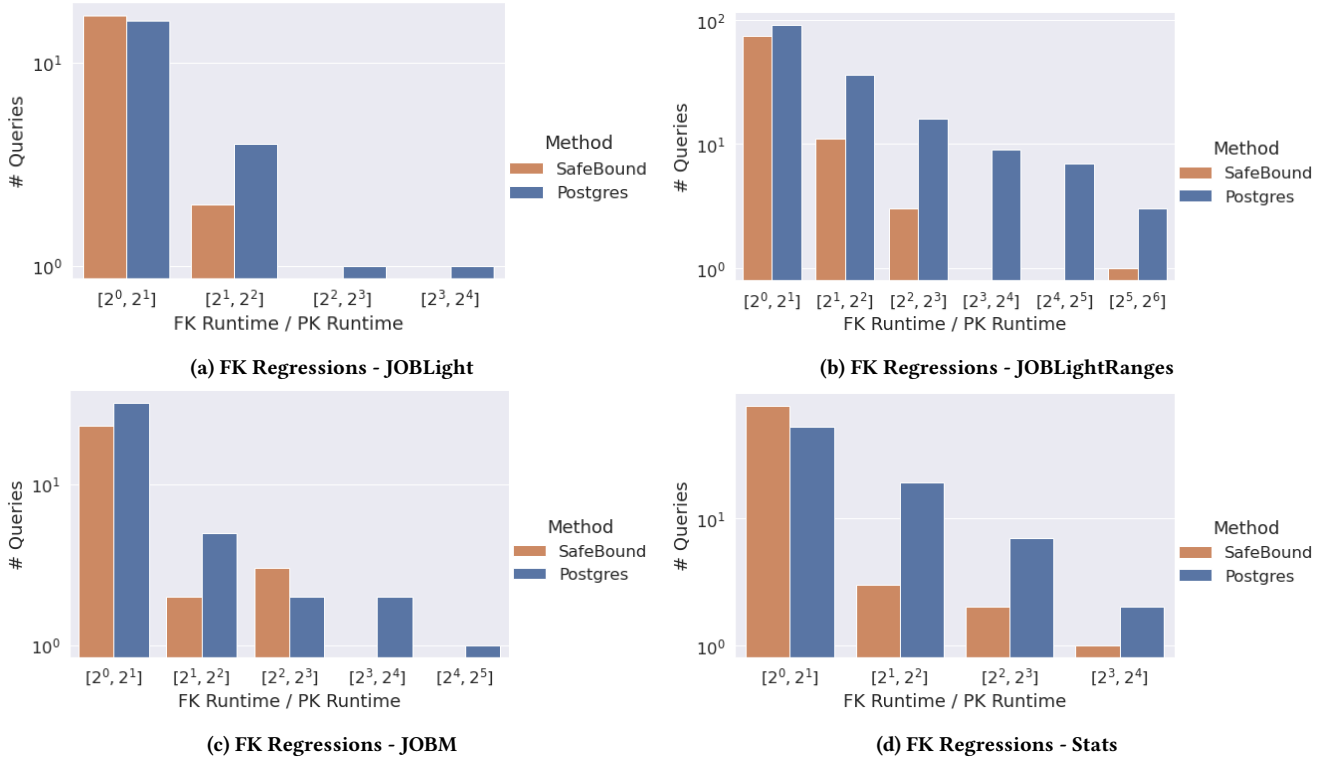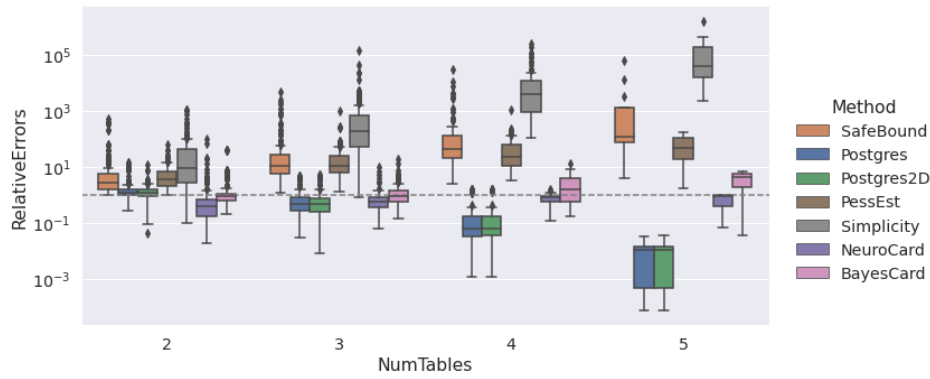
(a) FK Regressions - JOBLight

(b) FK Regressions - JOBLightRanges

(c) FK Regressions - JOBM
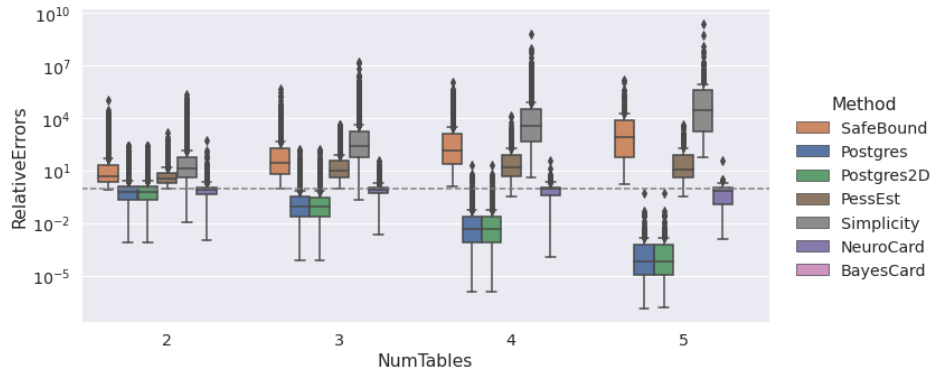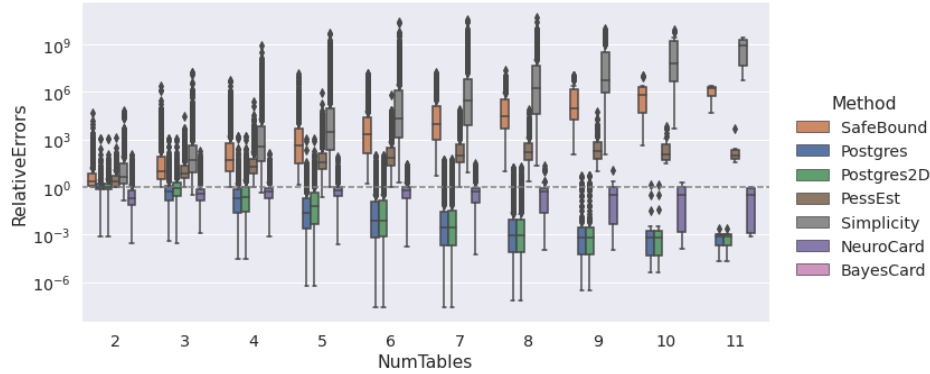

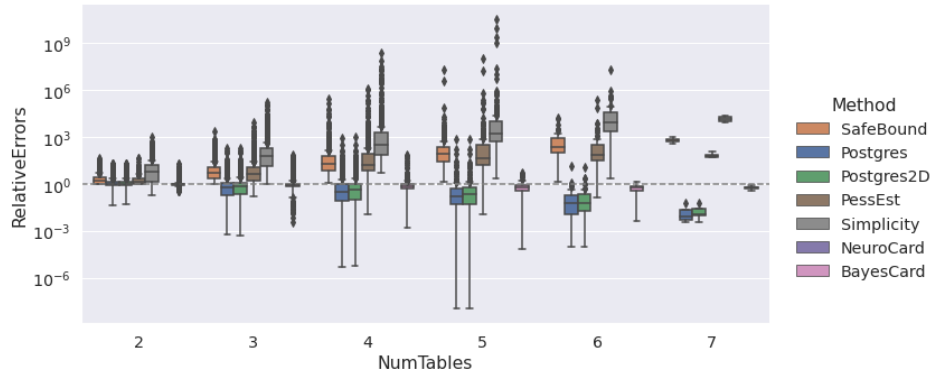(d) FK Regressions - Stats

Figure 3: FK Regressions Graphs

(a) Relative Error By Join Size - JOBLight

(b) Relative Error By Join Size - JOBLightRanges

(c) Relative Error By Join Size - JOBM

(d) Relative Error By Join Size - Stats

Figure 4: Relative Error By Join Size Graphs