Table 1: r decreases as the graph is corrupted.

Run	$p = 0$	$p = 2$	$p = 4$	$p = 6$
1	4.8E18	2.9E17	2.8E15	6.1E13
2	4.8E18	2.8E16	3.3E15	4.2E13
3	4.8E18	1.1E16	6.8E14	4.5E13

Table 2: Additional results on small datasets (RMSE).

Dataset	DGPG-L	SVR	RT	MLP
Weather	1.92	1.88	2.35	2.06
fMRI	0.020	0.020	0.021	0.024
ETEX	0.41	0.31	0.31	0.31

Usefulness of the Theorem: We perform a new experiment to show the usefulness of our theorem. It has been showed that $\alpha \gg \beta$ implies the algorithm benefits from graph information and thus converges faster. Now we use $r = \alpha/\beta$ to measure the richness of information of graphs: larger r suggests more informative graph structures. We corrupt the graph by randomly select $p = 0, 2, 4, 6$ edges and change them into erroneous connections. As p increases r becomes smaller (corrupted graph is less informative), which follows our analysis and expectation. Our theorem and the statistics it derives have potential applications, e.g. determining the network structures. We show 3 runs due to space limit.

We summarize the significance of the theorem as follows: 1) It presents a rigorous proof that considering graph structure can reduce sampling variances and thus encourage convergence; 2) Our additional experiment shows that the statistics developed in our proof have potential applications, e.g. determining more probable network structures; 3) Though the same goal can be achieved by observing ELBO in this example, they are derived from totally different mathematical principles. Their evaluation processes are also different: while evaluating ELBO requires stochastic sampling with the "re-parameterization trick", our statistics can be computed by algebraic manipulation and is more favorable; 4) One may further wander if it is possible to optimize α/β together with ELBO, which we leave for future investigation.

Additional Experiments on Small Datasets: We only present RMSE due to space limit. DGPG-L uses linear kernel (suggested by Reviewer #2); We further compare with standard regression algorithms (support vector regression, regression tree and MLP) where the output is a function of its neighbors' input (high-order graph information is lost).

Key Novelty and Contributions: 1) Though our method shares similarities with [11], optimizing ELBO with stochastic recursive sampling is a novel attempt and could potentially be extended to other structural domains (e.g. image). 2) Our theoretical analysis on the sampling variance is novel, introducing new insights and understandings. 3) Graph analysis community can better benefit from our work rather than from [11]: DGPG can model uncertainties and determine the importance of the connections with ARD kernels, which are important for graph analysis. 4) Our implementation is nontrivial: it transforms the graph into parallelizable data structures, and takes advantages of GPU acceleration. Both graph analysis and GP communities can benefit from our code, which will be available in our public repository.

Response to Reviewer #1: 1) We considered the evaluation of posterior variances in the experiment of Large Dataset (Section 4.3). Table 4 shows that our method can model uncertainty with satisfactory accuracy. 2) Units in Table 2 & 3 are different for each dataset. For instance the unit in Weather domain is Celsius. 3) An advantage of DGPG is that it does not require validation set. Train/valid/test splitting in the traffic flow prediction dataset is about 7/1/2, the same as [45] for comparison. DGPG can use the validation set for training as we discussed in Line 256-258. 4) Inducing points are optimized. 5) Number of iterations in the Weather/fMRI/ETEX/traffic datasets are 2000/5000/5000/10,000.

Response to Reviewer #2: 1) Usefulness of Theorem 1: Please see the discussions at the beginning. 2) DGPG with linear kernel: Please see the additional results in Table 2. 3) The number of inducing points is mostly chosen empirically. We would recommend to use $M = \sqrt{0.1N}$ where N is # of training instances, which is approximately the parameters we used. We find this value does not have too much effect as long as it is in a reasonable range. Initial locations of the inducing points are obtained by running k-means on the training data. 4) We thank the reviewer for pointing out the relevant work of [Bui] and [Burt et al.]. We will use the results from [Burt et al.] to enrich our discussion on complexity.

Response to Reviewer #3: 1) We hope that the reviewer's concern about novelty, contributions and comparison with stronger baselines can be addressed by our discussions and new experiments presented at the beginning. 2) [18], [19] and recent works of Bayesian graph ANN do not share the same goal of our paper, e.g. while DGPG learns from multiple signals some of their work consider 'node prediction' where only one signal is available and some nodes' values are missing. There is no obvious way to apply them to the datasets we consider. 3) GMRFs are a powerful model for interpolating signals over graphs. However our experiments require extrapolation across vertices (e.g. predict 90 nodes from 10). Preliminary results show that they are not suitable for comparison. We will review them in our discussion of related work. 4) In the new experiment the graph is no longer "ground true": we corrupt the graph with erroneous connections and analyze the results. 5) When the graph is fully connected, DGPG reduces to DGP. DGPG is a more generalized model, and considering the graph structure also brings novel analysis and insights.

Response to Reviewer #4: 1) Novelty & contributions: Please see the discussions at the beginning. 2) DGPG can be applied to directed graph so 'parent' is a more suitable term, e.g. in the traffic dataset the adjacent matrix is asymmetric. 3) In all current experiments $pa(v_k)$ include v_k. But it is still possible that the output signal does not depend on the input signal of the same node, so we leave it a task-specific choice whether the self-connection would be modeled. 4) We will describe the basic statistics that you mentioned in the table. In all the experiments each vertex has location information, and the graph is constructed according to Euclidean distances. We will explain this explicitly in our paper.

We thank all the reviewers for their insightful comments, which will be properly addressed in our later version.