First, thanks to the reviewers for the thorough and thoughtful reviews and comments!

Experiments: We stress (as we have tried to do in the paper) that our main contribution is theoretical. Our experiments are very preliminary. On the other hand, we were pleasantly surprised to find that our techniques, which we developed for the Erdös-Rényi (ER) model, have performance comparable to other algorithms (which have no theoretical guarantees) on real data. We include the experiments because they suggest that getting practical algorithms based on subgraph counts for real instances is a tenable (and in our opinion, interesting) direction for future research.

Responses to reviewer 1:   The reviewer said: "the assertion "We believe our techniques are of independent interest and applicability beyond the correlated ER model" [...] IMHO more work is needed to justify this assertion." We agree that more work is needed to pronounce that subgraph counting algorithms are successful beyond ER graphs; we did not intend to suggest otherwise! See "Experiments" above.

Regarding the relation to prior literature:

- For worst-case graphs the matching problem is NP-hard; thanks for pointing out the omission. One can reduce from Hamiltonian cycle, taking G_0 to be the target graph and taking G_1 to be a cycle on n vertices.
- The worst-case hypothesis testing problem was previously studied in the context of malware detection (see e.g. Section 2.2 of [Park-Reeves-Mulukutla-Sundaravel'10]). We are unaware of prior work on this question for random graphs.
- We will add a comparison between our work and that of Applebaum et al. and also Bhaskara et al. These papers do face some similar technical challenges having to do with subgraph counts in random graphs, but our goal is sufficiently different that it seems their techniques do not apply (at least not more naturally than what we already do).
- We should have mentioned the average-case graph isomorphism paper of Babai. The techniques there don't apply since the degree sequence (and in general neighborhood structure) cannot tolerate more than $o(1)$ noise.

The reviewer suggests that we "show that using simper family of black swans (cliques, cycles) does NOT work." We can show this in the relevant correlation regime (correlation $\gamma < 1 - \epsilon$). This because for any one subgraph H on v vertices, the variance in the subgraph count overwhelms the correlation by a multiplicative factor of $(1/\gamma)^{O(v)}$. To decrease the variance, we must take a family of size exponential in v (more precisely $(1/\gamma)^{O(v)}$). Since there are only $O(v)$ cycles and cliques on up to v vertices, there are not enough such "simple" graphs to reduce the variance.

Response to reviewer 2:   No concrete questions were posed. Thanks for the review!

Response to reviewer 3:   The recent work "Efficient random graph matching via degree profiles" by Ding et al. is *not* comparable with this work. Their algorithm tolerates significantly less noise: they require that two graphs differ on at most a $O(\frac{1}{\log^2 n})$ fraction of edges where n is the size of the base graph. In our paper we can tolerate a constant fraction of mismatched edges but still achieve exact recovery.

The reviewer says that the arXiv version is better-written and has more rigorous theorem statements. We agree; due to page limits we simplified. We will make the theorem statement more formal in the final version.

To compare our algorithm with GNN and SimGNN, we recast hypothesis testing as a classification problem. In the training phase, we generate pairs of graphs from the null hypothesis and the alternative hypothesis and feed them into neural network for supervised learning. Next we test the neural networks by asking them to classify newly generated pairs of graphs from both cases. The other papers mentioned by the reviewer are indeed relevant, and we will consider and incorporate comparisons into the final version of the paper.

For the suggestion about comparing our algorithm with those of by Feizi et al. and Dai et al.: these papers are clearly related to the problems we studied. The reason that spectral algorithms or canonical labeling algorithms don't work well is due to our high noise rate. For example, for two $\frac{1}{10}$-correlated graphs G_1 and G_2, the top eigenvector of G_1 and G_2 will be $\approx \frac{1}{10}$-correlated. In canonical labeling, the performance depends on the statistics used for labeling the vertices. If we label vertices just by degree statistics, there is no significant improvement can be made over the degree sequence approach. The reason our approach handles high noise is that we aggregate many statistics to reduce the variance.

The goal of the Facebook experiment is to give evidence that the subgraph count approach may work beyond the ER model, and can be useful for identifying the matching between two similar unlabeled "real-world" graphs. It is a simple algorithm and it achieves non-trivial number of correct matchings, even though the Facebook graph is dissimilar to a random graph. Greater improvements may be possible if the approach is developed further, and we find this an interesting direction for future research.