We thank the reviewers for their valuable positive feedback and comments. Please find our detailed answers below.

Reviewer 1

Links to other kernels [1] & using edge information: There is certainly a link to these kernels; the main difference is that these kernels exploit attribute and structural information by means of random walks (for label sequence generation), whereas our approach follows a WL subtree-based propagation scheme. Concerning the use of edge information, a straightforward extension for continuous high-dimensional edge attributes would apply WWL on the *dual* graph (where each edge is represented as a node, and connectivity is established if two edges in the primal graph share the same node), then combine it with the kernel on primal graphs via appropriate weighting. We will discuss the link to [1] in our revision and suggest extensions as future work.

Bibliography: Thank you, we checked the references and will update them for the revised version.

Reviewer 2

Information lost with WWL, weaknesses and issues, such as non unique embeddings and hashing: In the categorical case our kernel shares the propagation scheme of WL and WL-OA, which can lead to non-unique embeddings in rare cases. The key difference is that instead of comparing histograms, we compare distributions by means of optimal transport resulting in more granular similarities and thus higher classification performance. In the continuous case, to the best of our knowledge, WWL is the first WL-based method that does not rely on hashing; node attribute information is thus better exploited, as demonstrated by our empirical results.

Revise usage of KSVM: Thanks, we will clarify that we only use KSVM for the continuous variant of our kernel.

Assessment when comparing KSVM and SVM; price to pay for using indefinite kernels: For non-PSD kernels, KSVM has been proposed as a suitable replacement of classical SVM in machine learning. In practice, when the kernel is PSD, KSVM is equivalent to SVM, making the empirical and theoretical comparison of the two methods fair.[1] In our experiments with continuous attributes, we observed that the WWL kernel matrices are approximately PSD.

WL-OA with continuous node labels: WL-OA is based on node label histogram matching, hence it is not straightforward to extend it to the continuous case. While we agree that this would be interesting, the development of a continuous WL-OA variant is out of the scope of this work.

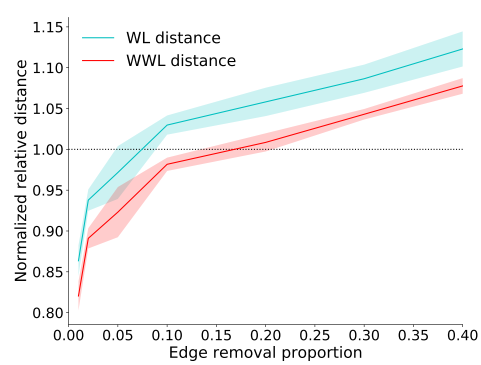
More graph kernel comparisons: Our experiments run all kernels on the same splits, coupled with a thorough hyperparameter selection, to guarantee a fair comparison. In an analogous setting, Kriege et al. (2016) showed, in the categorical case, that the shortest path and graphlet kernels perform worse than WL and WL-OA, so we did not include them. While we do not think such comparison is essential to our message, we could include it. For the continuous case, we provide an extensive comparison, including a variety of state-of-the-art kernels designed for continuously attributed graphs. We will describe the reasoning for our choice of comparison partners in more detail in the revised manuscript.

It should be "Reproducing" not "Reproducible". Thanks, we will correct this in the revised paper.

Reviewer 3

Gromov–Wasserstein: Thanks for the suggestion. Obtaining the Gromov-Wasserstein distance is computationally more demanding, but this direction is worth exploring in the future.

Isomorphism & distance 0: In the categorical case, for which WL can be applied, if the two graphs are isomorphic, the embeddings are guaranteed to be the same and the Wasserstein distance is 0. We are more interested in calculating the dissimilarity between graphs that are *not* isomorphic; here, WL and WWL will significantly differ, as WL is using a linear kernel between the histograms, whereas WWL characterizes differences in the distribution of labels.



SVM vs. k-NN & KSVM in WL setting: We had preliminary results with k-NN; following the literature & to ensure a fair comparison, we used SVM for the final experiments. KSVM is only used when PSD is not guaranteed; it is equivalent to SVM if the kernel is PSD[1], which is the case for WL, for example.

Improvements: Thanks for the interesting suggestion. We performed an additional experiment to evaluate the difference between WL and WWL for noisy E-R graphs ($n = 30$, $p = 0.2$). We report the relative distance between G and its permuted and perturbed variant G', w.r.t. a third independent graph G'' for an increasing noise level. We see that WWL is more robust against noise.

[1]Loosli, Gaëlle, Cheng Soon Ong, and Stéphane Canu. "Technical report: SVM in Krein spaces." (2013).