# Supplementary Materials for
# *A Persistent Weisfeiler–Lehman Procedure for Graph Classification*

**Bastian Rieck** [*1] **Christian Bock** [*1] **Karsten Borgwardt** [1]

## 1. Other Dissimilarity Measures

The proposed filtration in the paper is based on a metric, i.e. a distance, between multiset labels. As pointed out by one anonymous reviewer, though, it would also be possible to use other dissimilarity measures here. Specifically, since multisets essentially represent probability distributions, measures such as the *Kullback–Leibler divergence* (Kullback & Leibler, 1951) or the *Jensen–Shannon divergence* (Lin, 1991). Both of these measures can be easily used to generate weights and a subsequent filtration. More precisely, we define the following new label dissimilarity measure.

**Definition 1** (Label dissimilarity)**.** *Given two adjacent vertices $u$ and $v$ in a graph, let $\mathrm{l}_u^{(h)}$ and $\mathrm{l}_v^{(h)}$ refer to their multiset label in iteration $h$ of the Weisfeiler–Lehman stabilization procedure. Letting $\mathrm{d}_{\mathrm{KL}}$ denote the Kullback–Leibler divergence, we define their dissimilarity as*

$$\mathrm{d_L}^{\mathrm{KL}}(u,v) := \left[\mathrm{l}_u^{(h-1)} \neq \mathrm{l}_v^{(h-1)}\right] + \mathrm{d}_{\mathrm{KL}}\left(\mathrm{l}_u^{(h)}, \mathrm{l}_v^{(h)}\right) + \tau, \quad (1)$$

*where $[\cdot]$ is an Iverson bracket as in the main paper, i.e. a function that is $1$ whenever the condition in the bracket is satisfied, and $0$ otherwise, and $\tau \in \mathbb{R}_{>0}$ is an offset. Again, we use $\tau := 1$.*

Equivalently, it is possible to use the Jensen–Shannon divergence, i.e. $\mathrm{d}_{\mathrm{JS}}$, in Definition 1. If the symmetrised variant of these divergences is used, such that the divergence measure does not depend on the order in which edges have been specified, this leads to a filtration as specified in the main text. In contrast to many other filtrations in topological data analysis, this filtration is not induced by a metric. It is nonetheless valid and can be integrated in our framework. Table 1 depicts the results of training and testing with two divergence measures on the MUTAG data set, following the same experimental setup as in the paper. We observe that we do not reach the same level of accuracy as with the metric that was introduced in the main text. Please refer to our repository https://github.com/BorgwardtLab/P-WL for more experiments along these lines.

| | Kullback–Leibler | Jensen–Shannon |
|---|---|---|
| P-WL | $85.89 \pm 1.03$ | $85.71 \pm 0.67$ |
| P-WL-C | $89.03 \pm 1.95$ | $89.32 \pm 1.81$ |

*Table 1.* Classification accuracies on MUTAG when using divergence measures instead of metric proposed in the paper.

## 2. A Smoother Dissimilarity Measure

When designing a dissimilarity measure between connected vertices of a graph, another potential modification involves the comparison of the labels of the previous iteration step. In Equation 1, an Iverson bracket is used to compare them. However, given a dissimilarity measure $\mathrm{d_L}$, it would also be possible to calculate distances recursively, i.e. based on distances of a previous step. This would have a smoothing effect on the distances. Using again the Kullback–Leibler divergence, for example, one could set

$$\mathrm{d_L}^{(h)}(u,v) := \sum_h \mathrm{d}_{\mathrm{KL}}\left(\mathrm{l}_u^{(h)}, \mathrm{l}_v^{(h)}\right) + \tau, \quad (2)$$

where $\mathrm{d_L}^{(0)}(u,v)$ is a base distance that could be defined by checking the initial labels for equality, similar to the Iverson bracket in the main paper.

We use this distance and repeat the experiment described in Section 3.2 of the paper, which investigates the influence of the $h$ parameter. Figure 1 shows the behaviour of P-WL-C, which uses the metric described in the paper, and P-WL-CS, which uses the dissimilarity measure described in Equation 2; the additional "S" has been added to indicate that a *smooth* dissimilarity measure is employed. We observe that while the maximum mean accuracy obtained by P-WL-CS is slightly higher than that of P-WL-C, at least for a fixed $h$ parameter, the standard deviation tends to be a lot higher as well. For other data sets, such as MUTAG, we observe lower classification accuracies. Please refer to the repository for more experiments.

[*]Equal contribution [1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland. Correspondence to: Bastian Rieck <bastian.rieck@bsse.ethz.ch>, Karsten Borgwardt <karsten.borgwardt@bsse.ethz.ch>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

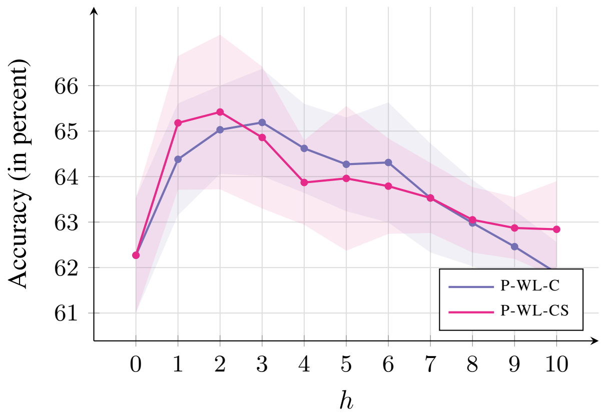


*Figure 1.* Performance of persistent subtree features based on either a regular or a smoothed variant of the Minkowski distance for the PTC-MR data set.

## References

Kullback, S. and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.

Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.