A Experimental setup

We describe the kernels under comparison, their parameters and the used datasets. All experiments were performed on an Intel Xeon E5-2690v4 machine at 2.6GHz with 384 GB of RAM.

A.1 Kernels

As a baseline we included the *node label kernel* (VL) and *edge label kernel* (EL), which are the dot products on node and edge label histograms, respectively, see [40, 25]. For the Weisfeiler-Leman subtree kernel (WL), the ℓ-step random walk kernel (RW) and the node-centric ℓ-walk kernel (NCW) and its variant with WL expressiveness (NCWWL) we chose the iteration number and walk length from $\{0, \ldots, 5\}$ by cross-validation. For RW, $\lambda_i = 1$ for $i \in \{0, \ldots, \ell\}$ was used. For NCW and NCWWL, we selected α from $\{0.01, 0.1, 1, 1000\}$ and β from $\{0, 0.5, 1\}$. We have not included extensions of the WL such as [23, 41], which could also be applied similarly to the node-centric ℓ-walk graph kernel. In addition we used a *graphlet kernel* (GL3) and the shortest-path kernel (SP) [4]. GL3 is based on connected subgraphs with three nodes taking labels into account similar to the approach used by Shervashidze et al. [39]. For SP we used the Dirac kernel to compare path lengths. We implemented the node-centric ℓ-walk graph kernel as well as all baselines in Java.[1] We performed classification experiments with the C-SVM implementation LIBSVM [6]. We report mean prediction accuracies and standard deviations obtained by 10-fold nested cross-validation repeated 10 times with random fold assignment. Within each fold all necessary parameters were selected by cross-validation based on the training set. This includes the regularization parameter C and kernel parameters.

A.2 Datasets

We tested on widely-used graph classification benchmarks datasets of the TUDATSETS repository [31] representing graphs from different domains. MUTAG, NCI1, NCI109 and PTC-FM represent small molecules, ENZYMES and PROTEINS are derived from macromolecules, and COLLAB and IMDBBIN are social networks. The datasets define binary graph classification experiments with exception of ENZYMES and COLLAB, which are divided into six and three classes, respectively. All graphs have node labels with exception of the social network graphs. We removed edge labels, if present, since they are not supported by all graph kernel implementations.

[1] Our code is publicly available at `http://anonymized`.