Skoltech

Skolkovo Institute of Science and Technology

Combinatorial and Neural Graph Vector Representations

Evgeny Burnaev

Head of ADASE group, Skoltech

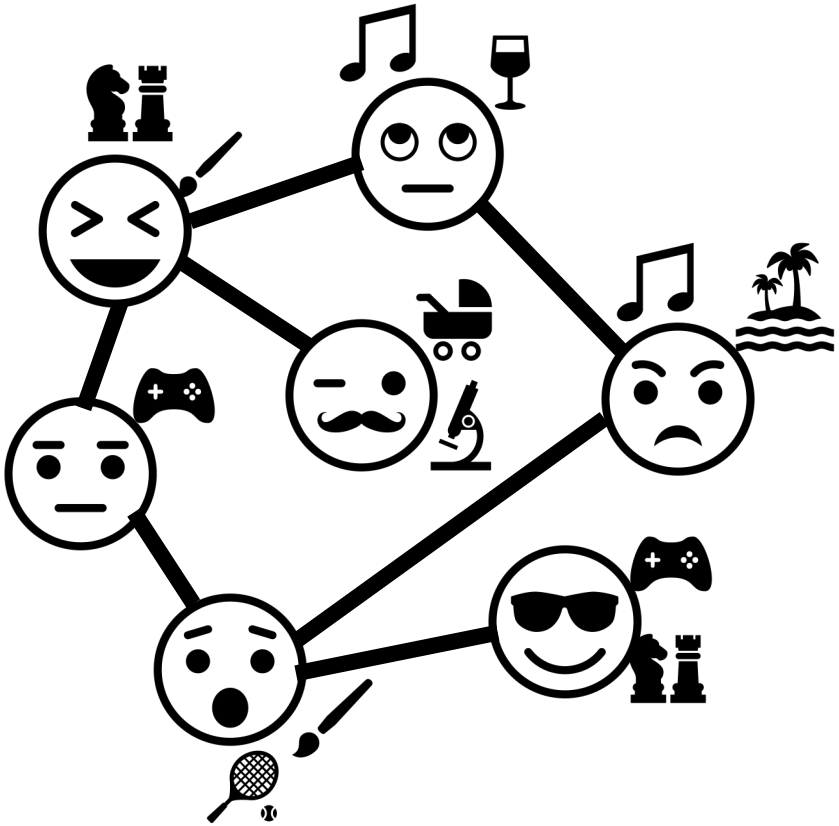
joint with *Sergey Ivanov, Sergey Sviridov*

Product recommendation

Skoltech
Skolkovo Institute of Science and Technology

How to find people that maximize product adoption?

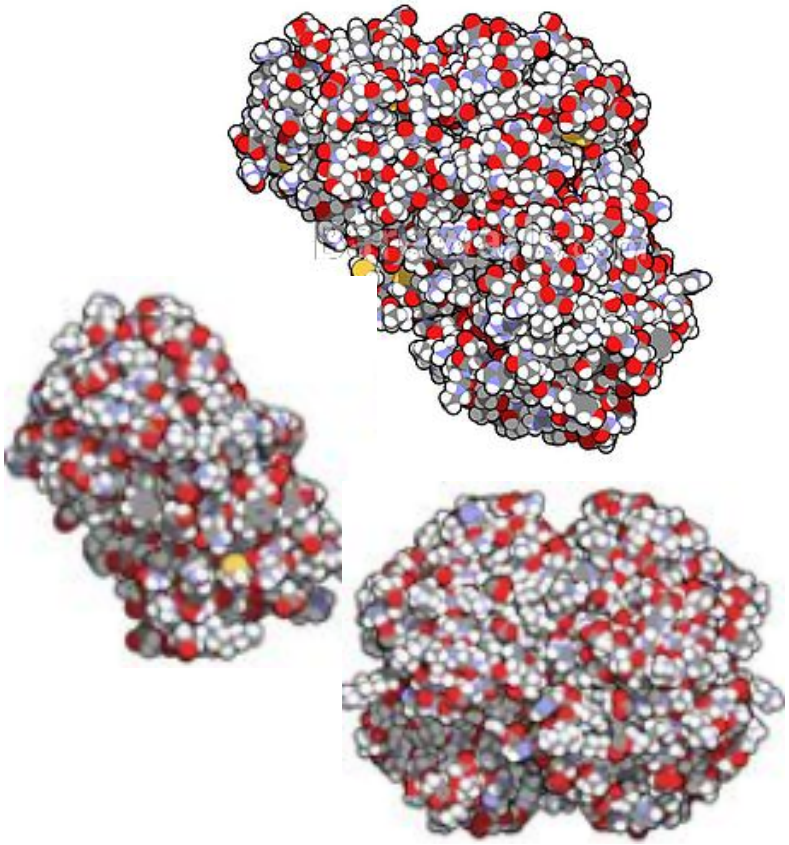
How to scale solutions to billions users and consider user preferences?



Protein function prediction

Skoltech
Skolkovo Institute of Science and Technology

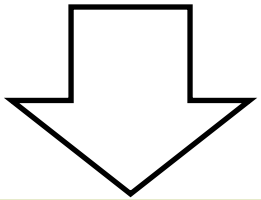
How to find similar proteins based on structural, physical, and chemical information?



What is common?

Skoltech
Skolkovo Institute of Science and Technology

- *We can model these problems with graphs;*
- *We can find solution for some classes of problems.*

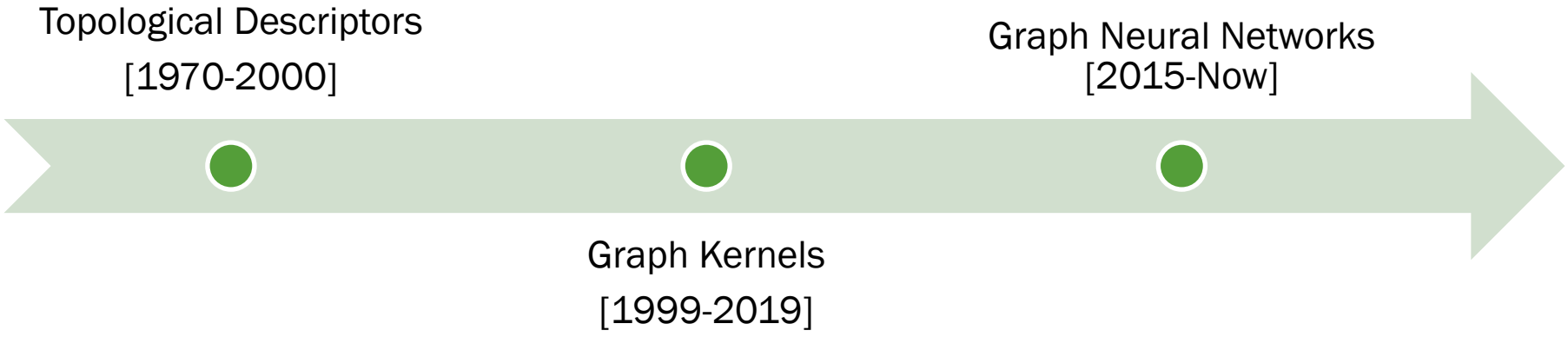


We can use ML methods on graphs to solve new instances.

How to represent graphs for ML models?

Representation learning on graphs

Skoltech
Skolkovo Institute of Science and Technology



Graph vector representation $v \in R^d$ is called *embedding*.

Topological descriptors

Skoltech
Skolkovo Institute of Science and Technology

Simple feature vectors or a scalar number (e.g. Wiener index).

Pros:

- Simple and inspired by chemical properties of studied networks

Cons:

- Very limited scope
- Ad-hoc design
- Prediction is not efficient (e.g. via knn)

[1] Handbook of molecular descriptors, Wiley & Sons 2008

Graph kernels

Skoltech
Skolkovo Institute of Science and Technology

Symmetric, positive semidefinite function that maps two graphs to a real number:

$$K(G_1, G_2) \mapsto R$$

Graph kernels: pros

Skoltech
Skolkovo Institute of Science and Technology

- More expressive than topological descriptors
- Includes different substructures
- Suitable for kernel machines (e.g. SVM)
- Non-linearity achieved by kernel trick
- Allows skipping explicit computation of vector representation
- Backed by theoretical developments in graph theory

Graph kernels: cons

Skoltech
Skolkovo Institute of Science and Technology

- Not scalable (often, $O(n^3)$ running time)
- Requires $N \times N$ kernel matrix
- Do not preserve graph isomorphism in feature space

Addressed in this work

[1] A survey on graph kernels, Kriege et al. 2019

Isomorphism property

Skoltech
Skolkovo Institute of Science and Technology

$$K(G_1, G_2) = < \varphi_1, \varphi_2 >,$$

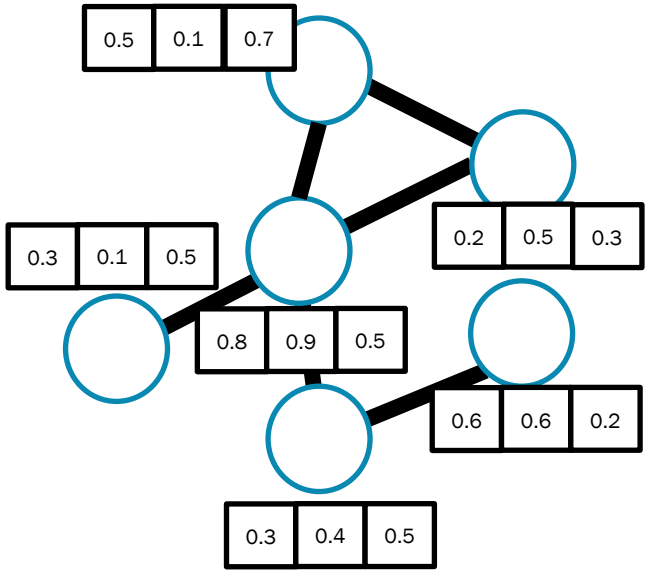
where $\varphi: G \mapsto R^d$

If φ is bijective, then we say graph kernel has *isomorphism property*. Such graph kernel minimizes loss of information.

Graph neural networks

Skoltech
Skolkovo Institute of Science and Technology

Graph embedding is initialized with random vector, which is updated to fit the given data (e.g. CNN-like models).



Graph neural networks: pros & cons

Skoltech
Skolkovo Institute of Science and Technology

Pros:

- SOTA performance in classification task
- Similar theoretical background

Cons:

- Complex models
- Hardly interpretable

Addressed in this work

[1] A Comprehensive Survey on Graph Neural Networks, Wu et al. 2019

Main goal

Skoltech
Skolkovo Institute of Science and Technology

To develop efficient graph representation that:

- Has isomorphism property
- Inherits strong graph kernel and neural network sides
- Test efficiency on real-world problems

Main results

Skoltech
Skolkovo Institute of Science and Technology

- Proposed and justified a new graph representation that **strictly provides isomorphism property**;
- Designed two approaches for **approximate efficient computation of embeddings**;
- Demonstrated **superior quality of embeddings** in graph classification problem.

1. Anonymous Walks

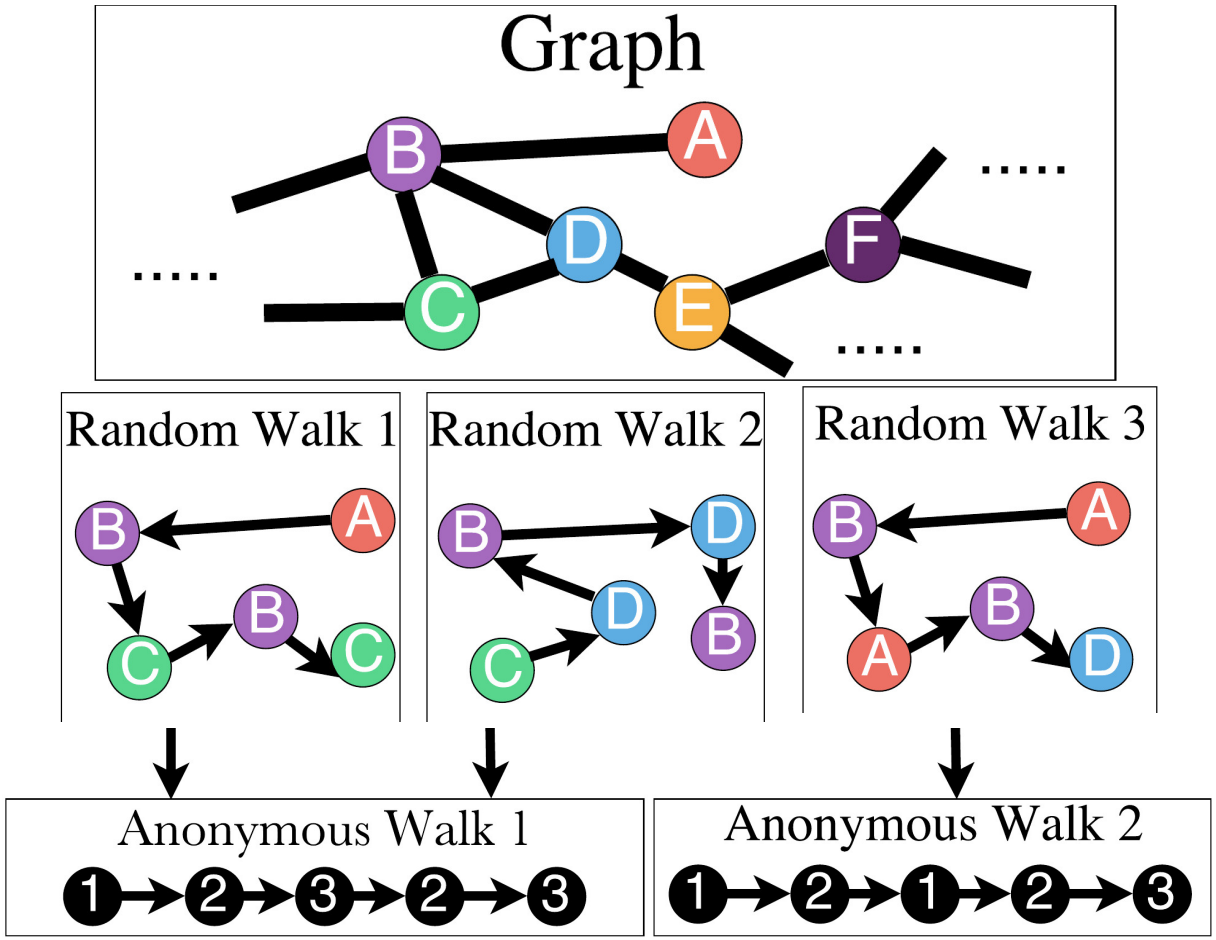
Anonymous Walks

Skoltech
Skolkovo Institute of Science and Technology

Definition:

If $w = (v_1, v_2, \dots, v_l)$ is a random walk then anonymous walk is the sequence $a = (a_1, a_2, \dots, a_l)$ where a_i is the first position of v_i in w.

Anonymous Walks



Reconstruction property

Theorem [Zhu & Micali, 2015]:

Let $B(v, r)$ be the induced graph at node v of radius r containing m edges, and D_l is a set of all possible anonymous walks of radius $l = 2(m + 1)$, that start at node v.

There is an algorithm to reconstruct a graph G that is isomorphic to $B(v, r)$.

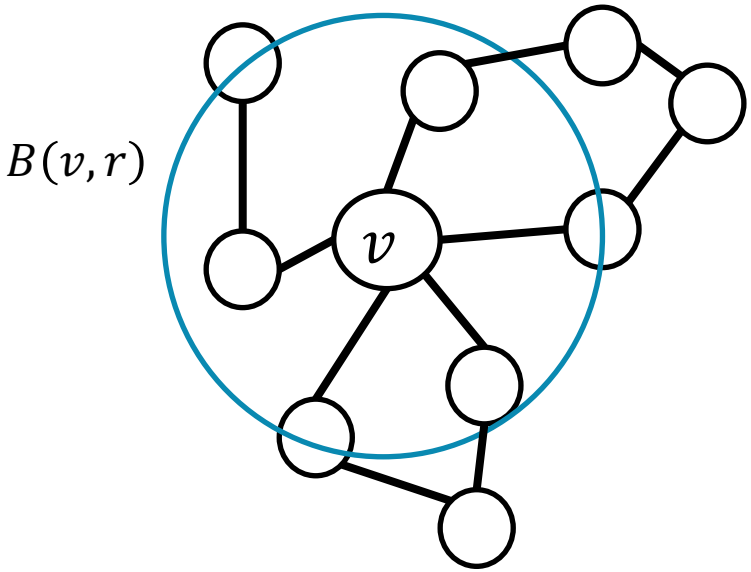
Theorem illustration

Radius $r = 2$

Edges $\mathrm{m} = 6$

Length $l = 2(m + 1) = 17$

Knowing distributions $D_2, D_3, \ldots, D_{17}$ we can obtain graph G that is isomorphic to $B(v, r)$.



$D_2 =$ { 0 1 0 ; 0 1 2 }

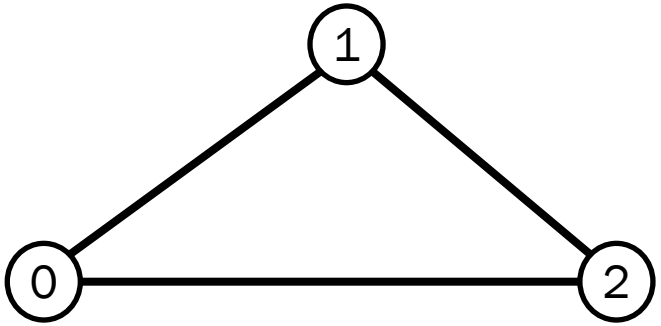
$D_3 =$ { 0 1 0 2 ; 0 1 0 1 ; 0 1 2 1 }

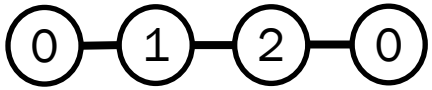
■ ■ ■

Covering walks

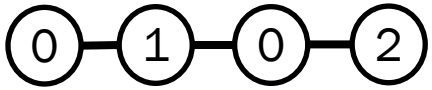
Definition:

If anonymous walk traverses each edge of the graph at least once, we call it *covering walk*.





covering walk



not covering walk

Isomorphism test

Theorem [This work]:

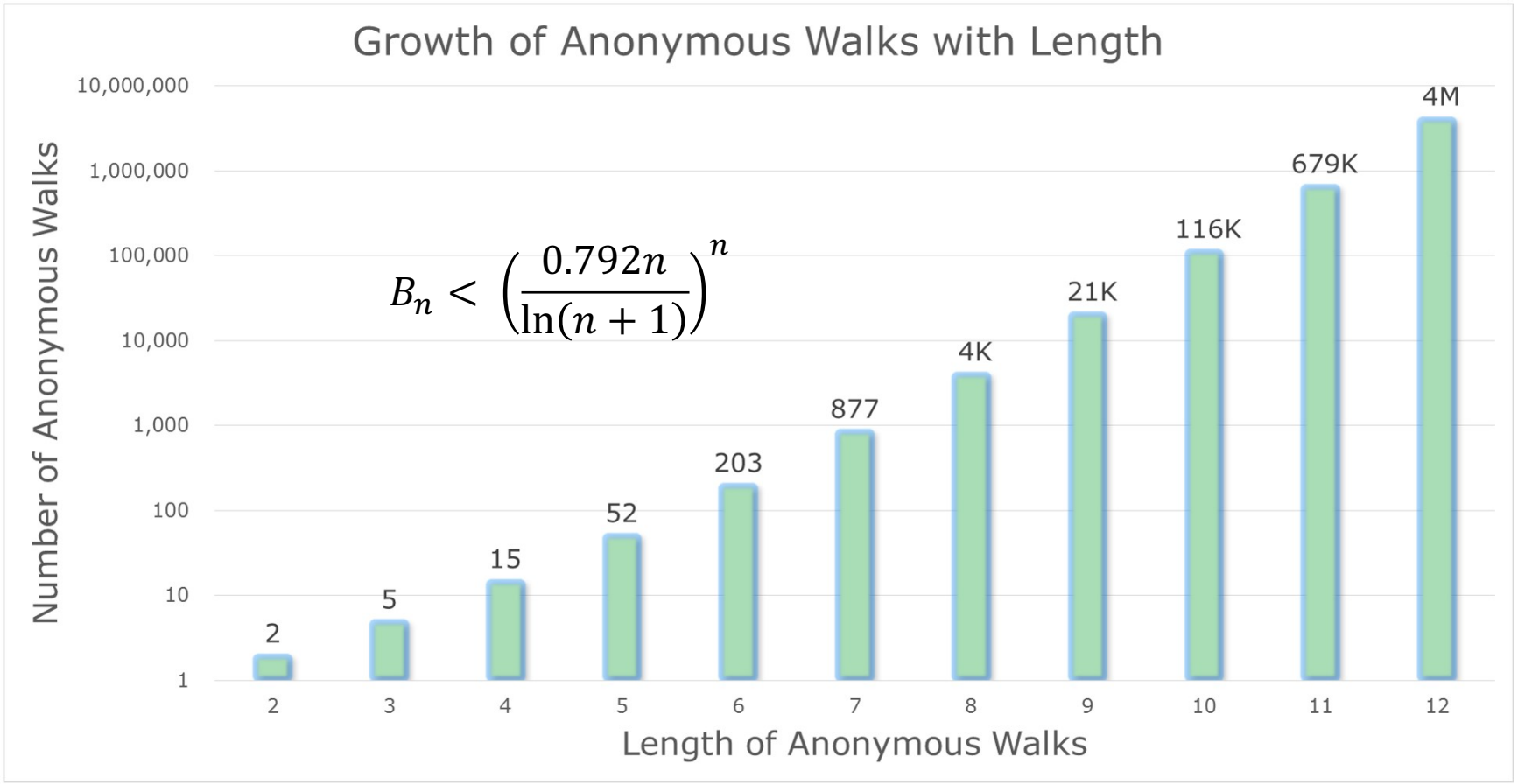
Let S_1 and S_2 be the sets of all covering walks of length $l = 2(m + 1)$ for graphs G_1 and G_2 with m edges. Two graphs are isomorphic if and only if $S_1 \cap S_2 \neq \emptyset$.

Running time complexity

Theorem [This work]:

The number of possible anonymous walks $|D_l|$ of length l in a graph is upper bounded by the Bell number B_{l-1}. The maximum for $|D_l|$ is attained for complete graph with l nodes.

$$|D_l| \leq B_{l-1}$$


Growth of Anonymous Walks with Length
Number of Anonymous Walks
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
$B_n < \left(\frac{0.792n}{\ln(n+1)}\right)^n$
2
5
15
52
203
877
4K
21K
116K
679K
4M
2
3
4
5
6
7
8
9
10
11
12
Length of Anonymous Walks

Combinatorial graph embeddings

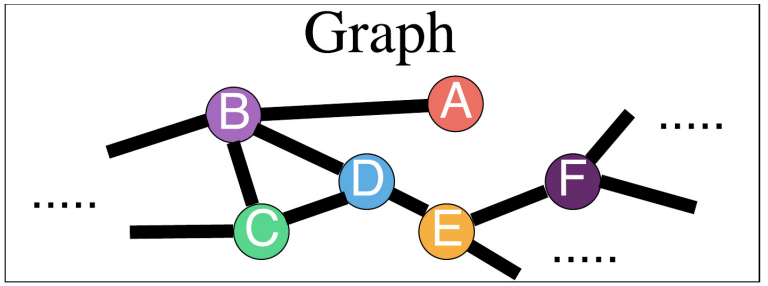
- $(a_1, a_2, \dots, a_\eta)$ all possible anonymous walks of length l.

l is a parameter for embeddings.

Combinatorial Graph Embedding

$\text{AWE}(G) = (p(a_1), p(a_2), \dots, p(a_\eta))$

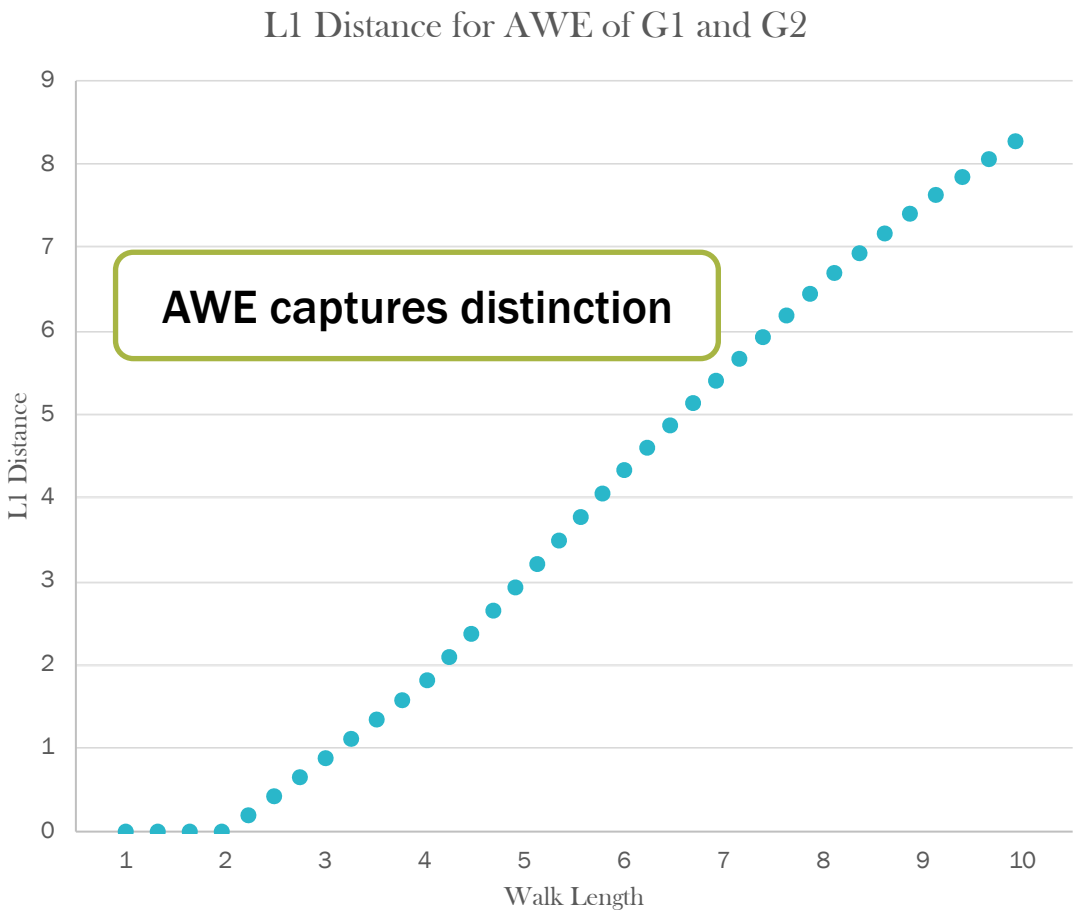
Where $p(a_i)$ is frequency of AW a_i



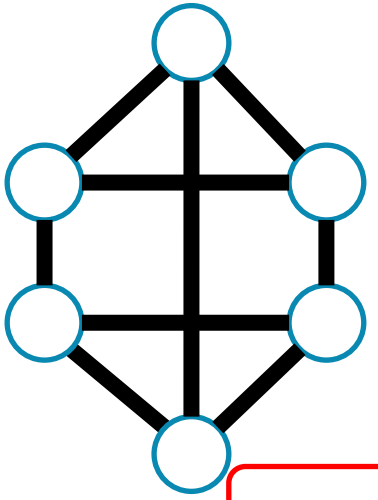
Graph ⇄ **Embedding**

$l = 2(m + 1)$

Example of Isomorphism Property

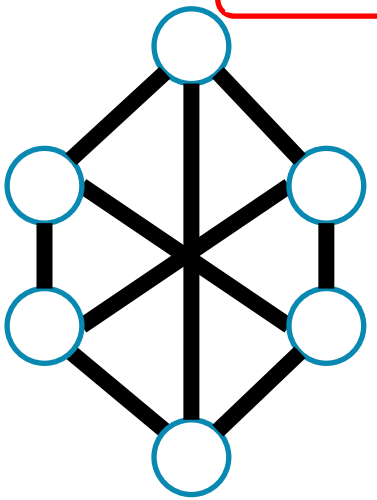


$$\|AWE(G)\|_1 = 1$$



G_1

WL fails to distinguish



G_2

Resolving computation complexity

Finding all possible anonymous walks $(a_1, a_2, \ldots, a_\eta)$ of length l can be expensive.

Instead, we can sample μ anonymous walks and compute embeddings from them.

Can we guarantee the quality of sampled embeddings?

Sampling inequality

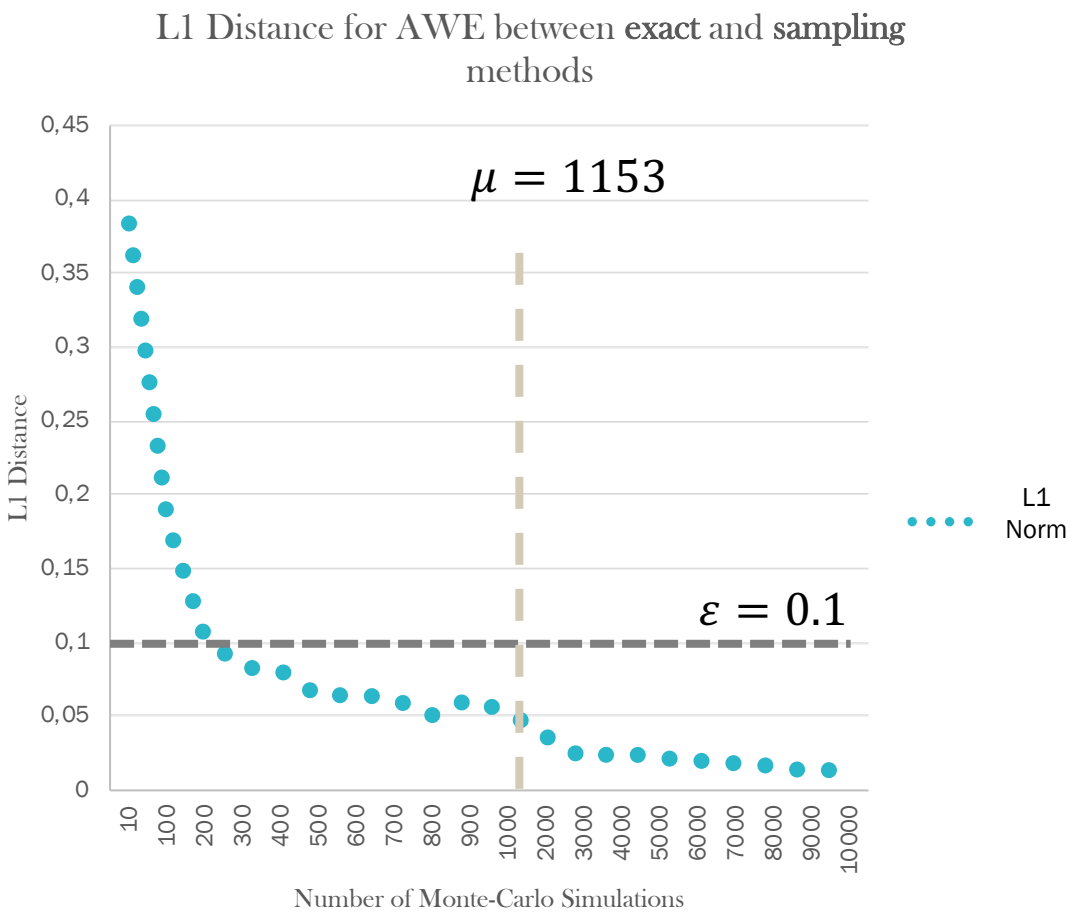
Theorem [This work]:

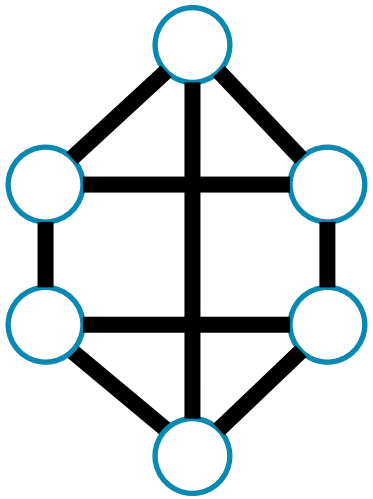
- D_l is *true* distribution of anonymous walks of length l in a graph G
- $\widehat{D}_l$ is *sampled* distribution of μ anonymous walks of length l from graph G.

For all $\varepsilon > 0$ and $\delta \in [0,1]$, the number of samples μ to satisfy $P\left(\left\|D_l - \widehat{D}_l\right\|_1 \geq \varepsilon\right) \leq \delta$ equals to:

$$\mu = \lceil \frac{2}{\varepsilon^2}\left(\log(2^\eta - 2) - \log(\delta)\right) \rceil$$

Example of Sampling Bound Property





G_1

$$\varepsilon = 0.1$$

$$\delta = 0.1$$

$$P\left(\left\|D_l - \widehat{D}_l\right\|_1 \geq 0.1\right) \leq 0.1$$

$$\mu = \left\lceil \frac{2}{\varepsilon^2}\left(\log(2^\eta - 2) - \log(\delta)\right)\right\rceil = 1153$$

Neural network embeddings

Sample a corpus of anonymous walks that start from the same node.

Initialize randomly graph embedding d and a matrix of embeddings W for each anonymous walk.

Maximize the average log probability of observing the corpus

Neural network embeddings

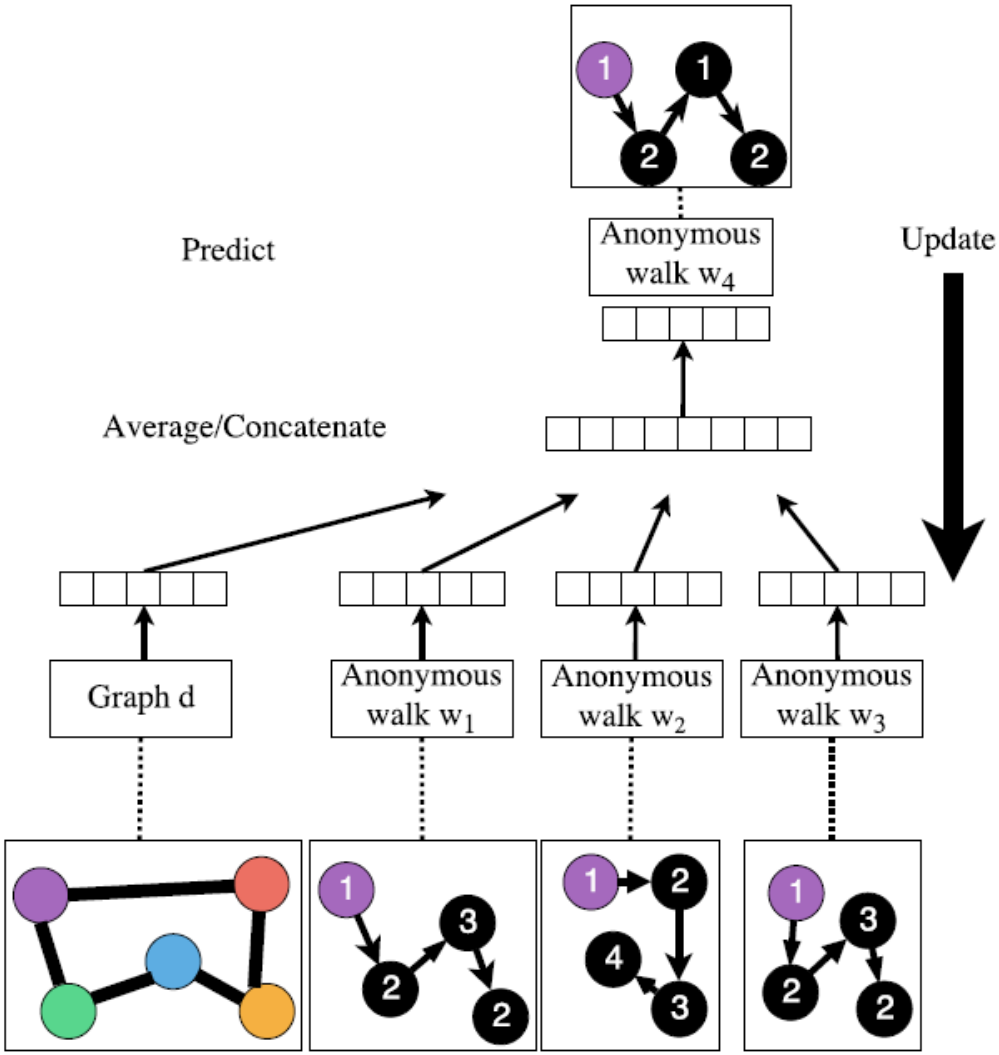
We optimize the objective

$$\frac{1}{T}\sum_{t=\Delta}^{T-\Delta} \log p(w_t|w_{t-\Delta}, \dots w_{t+\Delta}, d) \mapsto_{W,d} max$$

where

$$p(w_t|w_{t-\Delta}, \dots w_{t+\Delta}, d) = \frac{e^{y(w_t)}}{\sum_{i=1}^{\eta} e^{y(w_i)}}$$

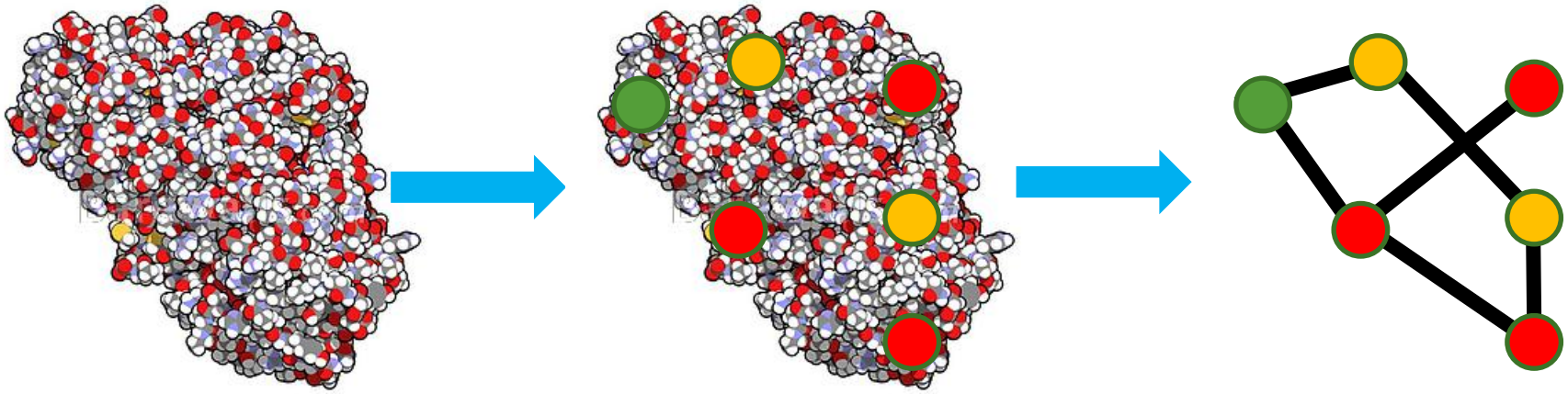
is the softmax probability of seeing anonymous walk in a graph and $y(w_i) = < w_i, [\frac{1}{2\Delta}\sum_{j=t+\Delta}^{t+\Delta} w_j ; d] >$ is similarity between walk w_i and its neighborhood.



2. Graph Classification

Protein function prediction

Skoltech
Skolkovo Institute of Science and Technology



Nodes: SSEs (helices, sheets, turns)

Edges: sequential or structural neighbors

Labels: length between atoms, polarity of SSE, etc. [1]

[1] Protein function prediction via graph kernels, Borgwardt et al. 2005

Graph classification problem

Given

- $T = \{(G_i, y_i)\}_1^N$ - train graph data set
- $Q = \{(G_i, y_i)\}_1^M$ - test graph data set

Using the train set T, find a function $f \in F = \{\phi: G \mapsto Y\}$ such that

$$acc = \frac{1}{M}\sum_{1}^{M}[f(G_i) = y_i] \mapsto max$$

Classification pipeline

Skoltech
Skolkovo Institute of Science and Technology

Prepare k-fold Cross-Validation splits.

⇩

Train model and choose the best hyperparameters for embeddings and classification models.

⇩

Evaluate the best model on test instances.

Embedding parameters

- Length of a walk
- Window size
- Embedding size

Model parameters

- Penalty term C
- Kernel type (e.g. Gaussian, Polynomial)
- Batch size

Datasets

	Dataset	Source	Graphs	Classes (Max)	Nodes Avg.	Edges Avg.
[1]	COLLAB	Social	5000	3 (2600)	74.49	4914.99
	IMDB-B	Social	1000	2 (500)	19.77	193.06
	IMDB-M	Social	1500	3 (500)	13	131.87
	RE-B	Social	2000	2 (1000)	429.61	995.50
	RE-M5K	Social	4999	5 (1000)	508.5	1189.74
	RE-M12K	Social	12000	11 (2592)	391.4	913.78
[2]	Enzymes	Bio	600	6 (100)	32.6	124.3
[3]	DD	Bio	1178	2 (691)	284.31	715.65
	Mutag	Bio	188	2 (125)	17.93	19.79

[1] Deep Graph Kernels, Yanardag et al. 2012.
[2] Protein function prediction via graph kernels, Borgwardt, 2005.
[3] Distinguishing enzyme structures from non-enzymes without alignments, Dobson & Doig, 2003

Skoltech
Skolkovo Institute of Science and Technology

Evaluation accuracy

	Algorithm	IMDB-M	IMDB-B	COLLAB
Neural	DGK	44.55 ± 0.52	66.96 ± 0.56	73.09 ± 0.25
Kernel	WL	49.33 ± 4.75	$\mathbf{73.4 \pm 4.63}$	$\mathbf{79.02 \pm 1.77}$
	GK	43.89 ± 0.38	65.87 ± 0.98	72.84 ± 0.28
	ER	OOM	64.00 ± 4.93	OOM
	kR	34.47 ± 2.42	45.8 ± 3.45	OOM
Ours	AWE (NN)	$\mathbf{51.54 \pm 3.61}$	$\mathbf{74.45 \pm 5.83}$	$\mathbf{73.93 \pm 1.94}$
	AWE (GK)	$\mathbf{51.58 \pm 4.66}$	73.13 ± 3.28	70.99 ± 1.49

[1] Deep Graph Kernels, Yanardag et al. 2012
[2] Weisfeiler-Lehman Graph Kernels, Shervashidze et al. 2011
[3] Efficient graphlet kernels for large graph comparison, Shervashidze et al. 2009
[4] Graph Kernels, Vishwanathan et al. 2010

Evaluation accuracy

	Algorithm	RE-B	RE-M5K	RE-M12K
Neural	DGK	78.04 ± 0.39	41.27 ± 0.18	32.22 ± 0.10
Kernel	WL	81.1 ± 1.9	49.44 ± 2.36	38.18 ± 1.3
	GK	65.87 ± 0.98	41.01 ± 0.17	31.82 ± 0.08
	ER	OOM	OOM	OOM
	kR	OOM	OOM	OOM
Ours	AWE (NN)	**87.89 ± 2.53**	**50.46 ± 1.91**	**39.20 ± 2.09**
	AWE (GK)	**82.97 ± 2.86**	**54.74 ± 2.93**	**41.51 ± 1.98**

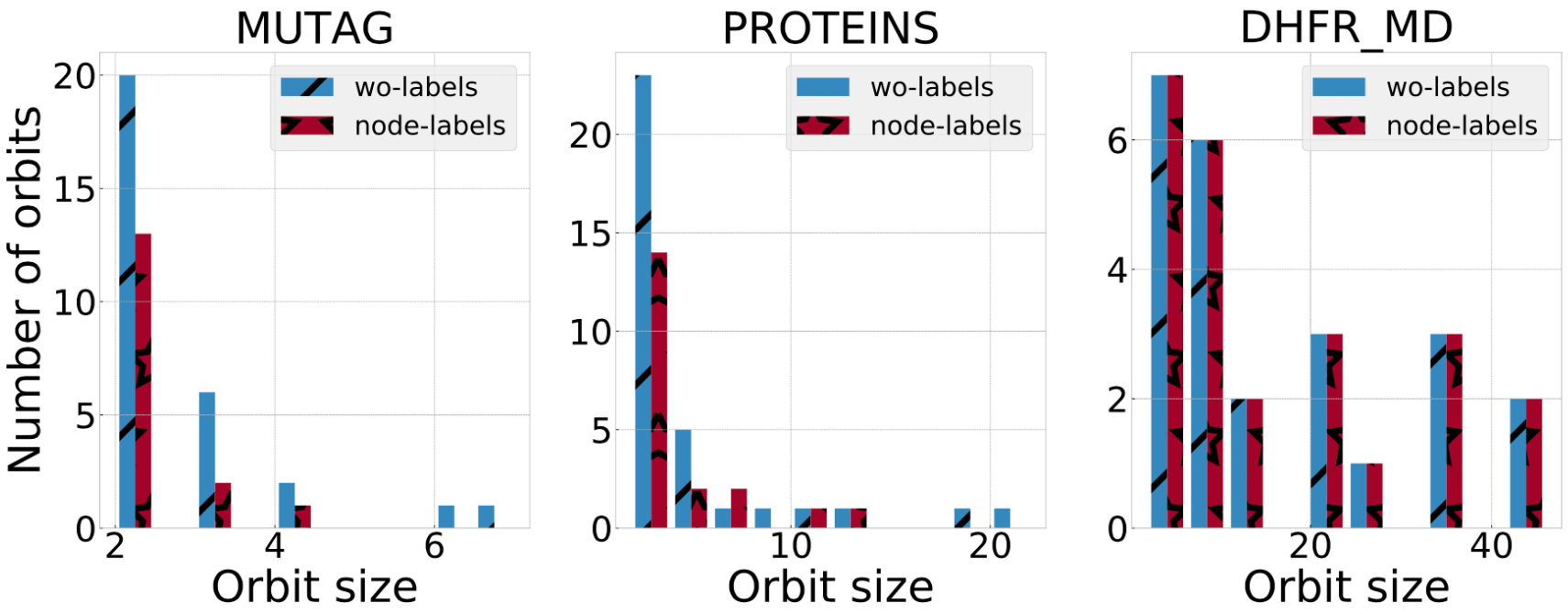
Evaluation accuracy

	Algorithm	Enzymes	DD	Mutag
Neural	DGK	27.08 ± 0.79	–	82.66 ± 1.45
Kernel	WL	**53.15 ± 1.14**	**77.95 ± 0.70**	80.72 ± 3.00
	GK	32.70 ± 1.20	**78.45 ± 0.26**	81.58 ± 2.11
	ER	14.97 ± 0.28	OOM	71.89 ± 0.66
	kR	30.01 ± 1.01	OOM	80.05 ± 1.64
Ours	AWE (GK)	**35.77 ± 5.93**	71.51 ± 4.02	**87.87 ± 9.76**

3. Learning to correct predictions for Graph Classification with noisy labels

Graph classification and graph isomorphism

- Even theoretically expressive graph classification models do not fully exploit the isomorphism between the graphs in the data set at hand
- Most of the data sets used in graph classification have isomorphic instances
- In some datasets the fraction of the unique non-repeating graphs is as low as 20% of the total size, while those graphs that are isomorphic often have non-unique target label


MUTAG
PROTEINS
DHFR_MD
wo-labels
node-labels
Number of orbits
Orbit size
Orbit size
Orbit size

data set	Size, N	Isomorphic, %	Noise, %
Letter-low	2250	99.78	8.72
IMDB-MULTI	1500	80.8	6.39
AIDS	2000	62.95	0.13
COX2	467	60.6	0.6
IMDB-BINARY	1000	57.9	0.67
MUTAG	188	42.02	0.49
COLLAB	5000	21.54	0.11
PROTEINS	1113	13.57	0.1
NCI1	4110	12.73	0.01
NCI109	4127	12.58	0.01
ENZYMES	600	1.67	0
REDDIT-BINARY	2000	0.2	0
REDDIT-MULTI-12K	11929	0.14	0

Table 1: The number of isomorphic graphs and noisy target labels in popular graph data sets. The last column indicates the proportion of graphs that contain at least one isomorphic copy in the same data set, which has a different target label.

Algorithm 1 Error-Correcting Graph Classification

Get predictions of the model on Y_{test};
# Error-Correcting Phase
for $G_i \in Y_{test}$ **do**
    Initialize orbit $O_i = [G_i]$
    for $G_j \in Y_{train}$ **do**
        if $G_i \cong G_j$ **then**
            Add G_j to O_i
        end if
    end for
    Predict $y_i = \arg\max_y \mathbb{P}(y|G_i, O_i)$.
end for

Error-correction performance

Graph Isomorphism Network (GIN)

Weisfeiler-Lehman graph kernel (WL)

Vertex histogram kernel (V)

- Note that GIN and WL kernels have sufficient expressiveness to recognize most of the isomorphic graphs in real data sets and hence one may hypothesize that the error-correcting phase is unnecessary for these methods.
- This hypothesize is false and they *do benefit* from additional correction of the predictions
- We add a prefix EC (error correction) for each model that follows Algorithm 1 and a prefix ECU that has an Error-Correcting phase only for the orbits with a Unique target label

Table 2: Mean classification accuracy for test sets Y_{test} and Y_{iso} (in brackets) in 10-fold cross-validation. Top-1 result for each type of algorithm in bold. Error correction consistently achieves increase of accuracy for each model.

	MUTAG	IMDB-B	IMDB-M	COX2	AIDS	PROTEINS
GIN	0.829 (0.840)	0.737 (0.733)	0.501 (0.488)	0.82 (0.872)	0.996 (0.998)	0.737 (0.834)
GIN-EC	0.856 (0.847)	0.737 (0.731)	0.499 (0.486)	0.795 (0.83)	0.996 (0.999)	0.729 (0.709)
GIN-ECU	**0.867** (1.000)	**0.756** (1.000)	**0.522** (1.000)	**0.838** (1.000)	**0.996** (1.000)	**0.742** (1.000)
WL	0.862 (0.867)	0.734 (0.990)	0.502 (0.953)	0.800 (0.974)	0.993 (0.999)	0.747 (0.950)
WL-EC	0.870 (0.838)	0.724 (0.715)	0.495 (0.487)	0.794 (0.844)	0.994 (0.999)	0.740 (0.742)
WL-ECU	**0.907** (1.000)	**0.736** (1.000)	**0.504** (1.000)	**0.810** (1.000)	**0.994** (1.000)	**0.749** (1.000)
V	0.836 (0.902)	0.707 (0.820)	0.503 (0.732)	0.781 (0.966)	0.994 (0.997)	0.726 (0.946)
V-EC	0.827 (0.844)	0.724 (0.728)	0.496 (0.481)	0.768 (0.852)	0.996 (0.999)	0.719 (0.741)
V-ECU	**0.859** (1.000)	**0.750** (1.000)	**0.517** (1.000)	**0.794** (1.000)	**0.996** (1.000)	**0.729** (1.000)

Table 3: Mean classification accuracy for test sets Y_{test} in 10-fold cross-validation on data sets with no noise. Each cell is $a \pm b(c)$, where a is mean accuracy, b is standard deviation, and c is absolute difference of accuracy between original (noisy) and new data sets.

Data	DGCNN	DiffPool	GIN	Baseline
MUTAG	80.59±11 (-1.60)	81.48±8 (-1.75)	**84.93±9** (-0.55)	71.08±6 (5.84)
IMDB-B	70.53±6 (-1.23)	72.96±5 (-2.43)	**74.84±3** (-4.84)	52.47±5 (0.83)
IMDB-M	46.11±4 (-0.17)	45.90±6 (-0.28)	**46.82±6** (-0.93)	41.11±6 (-7.07)
COX2	69.62±7 (9.05)	**75.11±9** (7.19)	73.83±5 (6.92)	68.30±11 (8.31)
BZR	79.33±6 (0.72)	**81.64±7** (1.48)	80.69±5 (3.28)	78.43±7 (4.12)
PTC_FM	**63.96±5** (-3.41)	59.46±6 (0.37)	57.76±6 (2.50)	60.89±6 (-2.43)
PTC_MM	**66.83±3** (-4.23)	59.97±9 (5.00)	60.36±6 (4.71)	58.66±7 (2.66)
PTC_MR	**60.72±4** (-4.48)	57.69±5 (-0.98)	56.01±8 (-0.44)	59.72±7 (-3.17)
PTC_FR	**66.01±2** (-2.76)	63.51±2 (0.30)	61.39±6 (4.61)	60.86±9 (1.73)

4. fMRI-based Depression Diagnostics

fMRI-based Depression Diagnostics

- 50 million people worldwide have epilepsy, making it one of the most common neurological diseases;
- 300 million people globally affected with depression as common psychiatric disorder;
- 37% of **epilepsy** patients suffer from **depression**.

The biomarkers of major depressive disorder as the **concomitant disease** are still unclear as well as their relation to epilepsy.

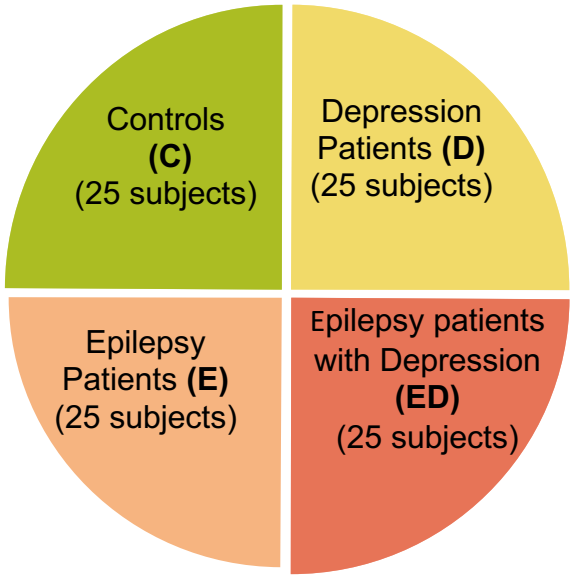
World Health Organization. (2018, May 21). Retrieved from http://www.who.int/en/news-room/fact-sheets/detail
Kwon O-Y, Park S-P. Depression and Anxiety in People with Epilepsy. Journal of Clinical Neurology (Seoul, Korea). 2014;10(3):175-188. doi:10.3988/jcn.2014.10.3.175.

Data set

Skoltech
Skolkovo Institute of Science and Technology

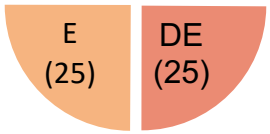
Medical partner Research and Clinical Center of Neuropsychiatry named after Solovyov.

Dataset of **100 functional 1.5 T MRI** sets in four groups (mean age: 33± 9 years).

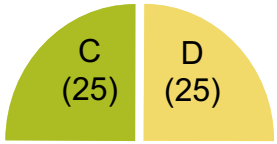


Classification tasks:

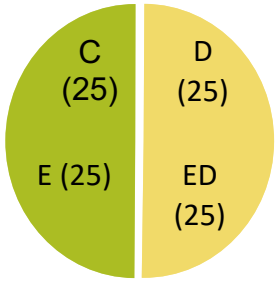
E vs DE:



D vs C:

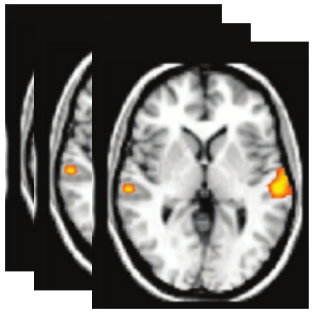


D vs NoD:



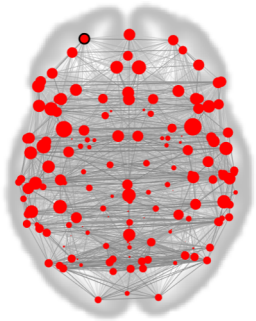
Data content

Skoltech
Skolkovo Institute of Science and Technology



Resting state T2* EPI fMRI
[64, 64, 30] voxels
133 time steps

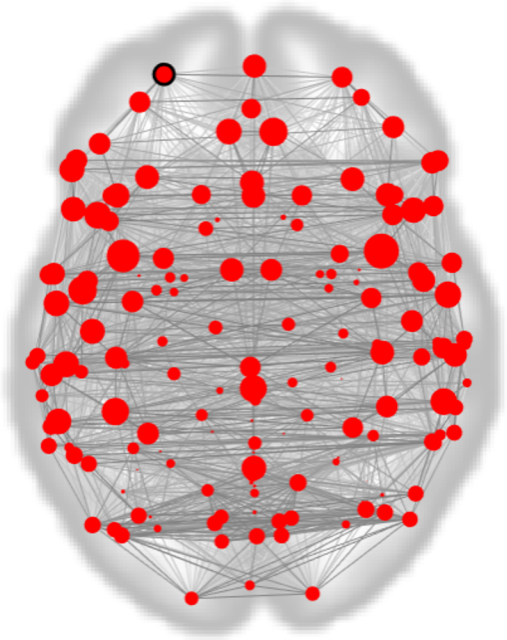
Preprocessing:
Time series extraction



- 117 regions (Automated Anatomical Labeling atlas)
- Time series for each region
- Correlation matrix

Connectivity matrix

Skoltech
Skolkovo Institute of Science and Technology



	ROI_1	ROI_2	...	ROI_116	ROI_117
ROI_1	1	0.2418	...	0.034	0.086
ROI_2	0.2418	1	...	0.053	0.247
...	...	....	1	...	...
ROI_116	0.034	0.053	...	1	0.017
ROI_117	0.0.086	0.247	...	0.017	1

Feature-based graph classification

Skoltech
Skolkovo Institute of Science and Technology

1. Feature based methods

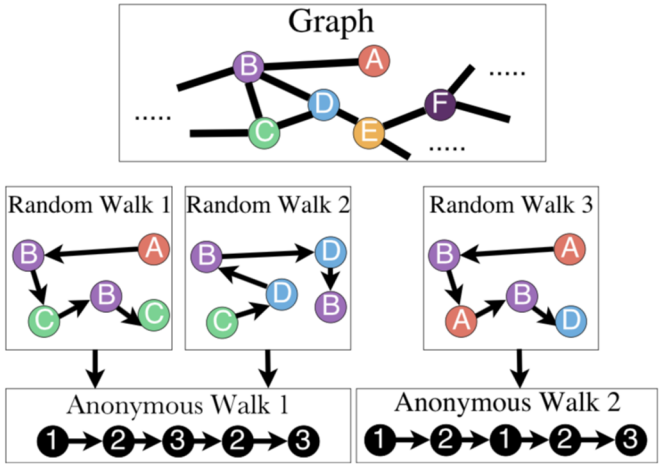
- Degree
- Average neighbor degree
- Degree centrality
- Betweenness/ closeness centrality
- Clustering coefficient
- Average path length
- Local/ global efficiency

$$E(G) = \frac{1}{n(n-1)} \sum_{i<j \in G} \frac{1}{d(i,j)}$$

where n is number of nodes - regions in a graph and $d(i,j)$ is the length of the shortest path between nodes i and j.

2. Kernel methods

- Weisfeiler-Lehman kernel (WL)
- **Random walks (Anonymous walk embedding (AWE))**



Evaluation

- 4 groups of patients: healthy (H), depression (D), epilepsy (E), depression + epilepsy (DE)
- Each group has 25 graphs
- Two classification tasks: DvsH and DvsDE

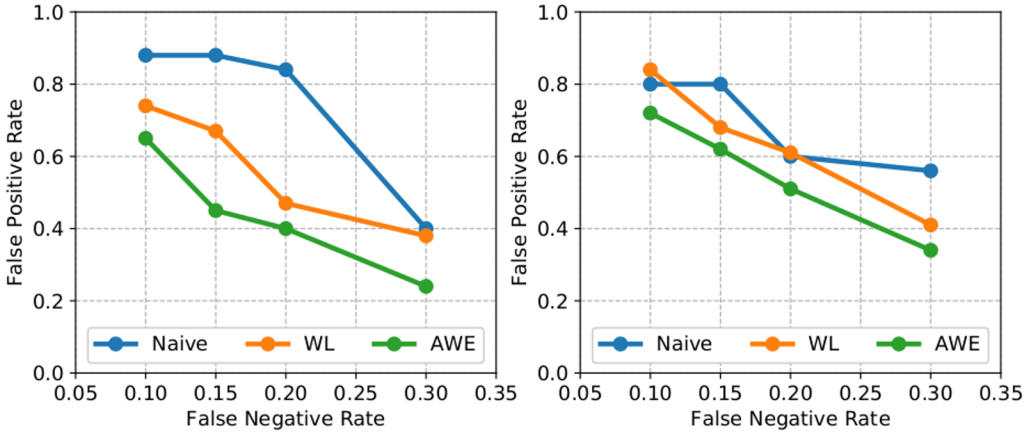


Fig.1. Classification results on DvsH and EvsDE tasks in terms of False Negative Rate vs. False Positive Rate performance curves. Left image: DvsH task. Right image: EvsDE task.

Task	Naïve	WL	AWE
DvsH	$73 \pm 15\%$	$78 \pm 15\%$	$\mathbf{80 \pm 12\%}$
EvsDE	$67 \pm 15\%$	$75 \pm 14\%$	$\mathbf{76 \pm 16\%}$

Main results

Skoltech
Skolkovo Institute of Science and Technology

- Proposed and justified a new graph representation that **strictly provides isomorphism property**;
- Designed two approaches for **approximate efficient computation of embeddings**;
- Demonstrated **superior quality of embeddings** in graph classification problem.

Published Papers

Skoltech
Skolkovo Institute of Science and Technology

- S. IVANOV, E. BURNAEV, *"ANONYMOUS WALK EMBEDDINGS"*, **PROCEEDINGS OF INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML),** 2018
- S. IVANOV, N. DURASOV, E. BURNAEV, *"LEARNING NODE EMBEDDINGS FOR INFLUENCE SET COMPLETION"*, **IEEE INTERNATIONAL CONFERENCE IN DATA MINING (ICDM) 2018 WORKSHOPS PROCEEDINGS,** 2018
- M. SHARAEV, A. ARTEMOV, A. BERNSTEIN, E. KONDRATYEVA, S. SUSHCHINSKAYA, E. BURNAEV, S. IVANOV, "LEARNING CONNECTIVITY PATTERNS VIA GRAPH KERNELS FOR FMRI-BASED DEPRESSION DIAGNOSTICS", **IEEE INTERNATIONAL CONFERENCE IN DATA MINING (ICDM) 2018 WORKSHOPS PROCEEDINGS,** 2018
- S. IVANOV, S. SVIRIDOV, E. BURNAEV, "UNDERSTANDING ISOMORPHISM BIAS IN GRAPH DATA SETS", **SUBMITTED, ARXIV 1910.12091**, 2019