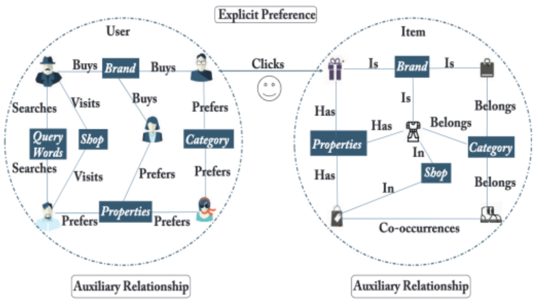
Sparse Representation Method for Whole Graph Embedding

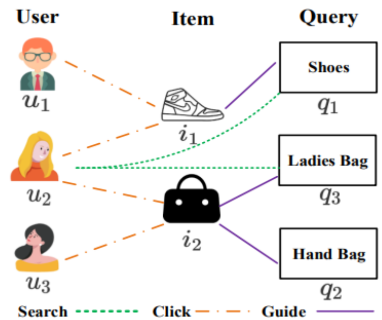
Oral Comprehensive Exam
Kaveen Liyanage

Dr. Bradley Whitaker
Dr. Rob Maher

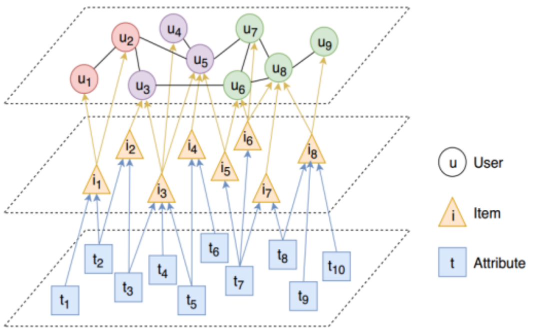
Graphs are helpful



(a) E-commerce recommendation HG [20].

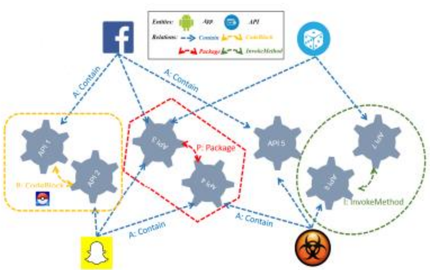


(b) Intent recommendation HG [19].

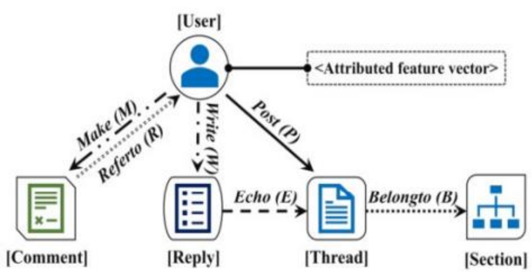


(c) User Profiling HG [126].

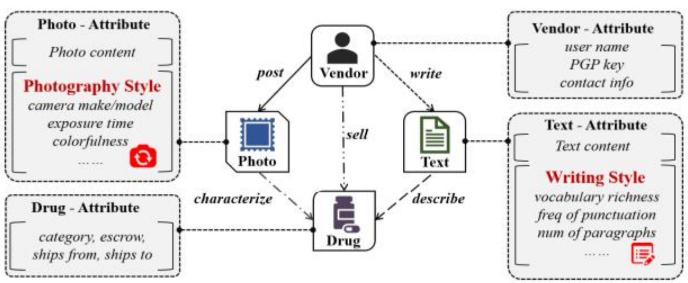
Fig. 3: The representative HGs in E-commerce.



(a) Malware detection [7].



(b) Key player identification [99].



(c) Drug trafficker identification [101].

Fig. 4: The representative HGs in cybersecurity applications.

1. Wang, X., Bo, D., Shi, C., Fan, S., Ye, Y. and Philip, S.Y., 2022. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. *IEEE Transactions on Big Data*.

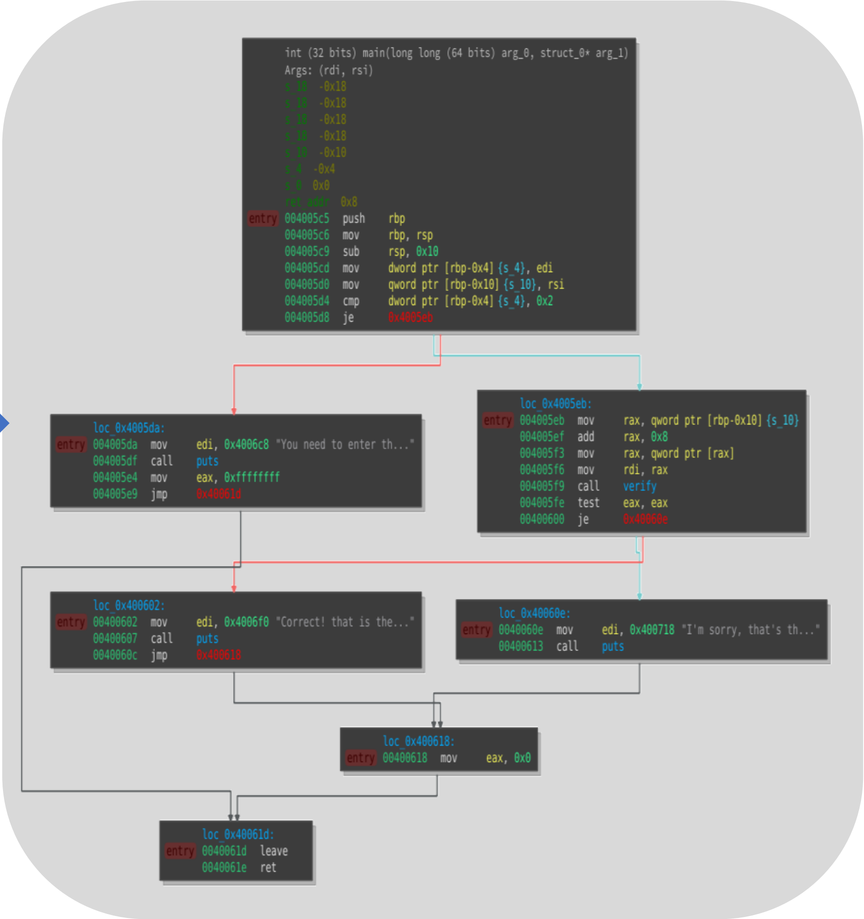
Overview

- Introduction
  - Motivation / Problem
  - Goal
  - Objectives
- Background
  - Graph embedding
  - Sparse representation
  - Feature ranking
- Workflow
  - Preliminary work
  - Proposed timeline
  - Future directions

Malware detection by Control Flow Graphs (CFG) of binary files

```
402000 endbr64
402004 sub    rsp, 0x8
402008 mov    rax, qword ptr
[0x409fd0]
40200f test   rax, rax
402012 je     0x402016
402016 add    rsp, 0x8
40201a ret
402014 call   rax
402020 push   qword ptr [0x409e70]
402026 bnd jmp qword ptr [0x409e78]
40202d nop    dword ptr [rax]
402030 endbr64
402034 push   0x0
402039 bnd jmp sub_402020
40203f nop
402040 endbr64
402044 push   0x1
402049 bnd jmp sub_402020
40204f nop
402050 endbr64
402054 push   0x2
402059 bnd jmp sub_402020
40205f nop
402060 endbr64
402064 push   0x3
402069 bnd jmp sub_402020
40206f nop
402070 endbr64
402074 push   0x4
402079 bnd jmp sub_402020
40207f nop
402080 endbr64
402084 push   0x5
```

Python
angr library

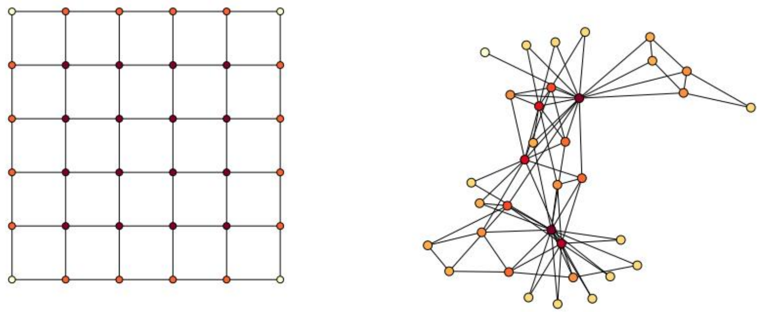


Binary File

Control Flow Graph (CFG)

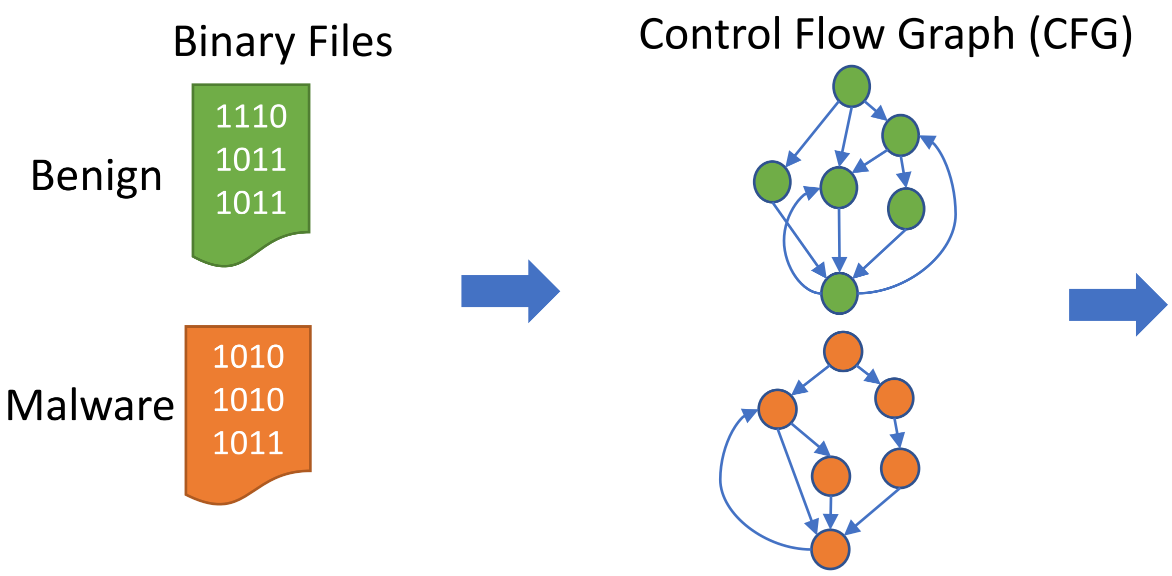
Funded by the US Department of Homeland Security under Contract No. 70RSAT22CB0000005

Graph embedding is an important process



(a) Grid (Euclidean). (b) Arbitrary graph (Non-Euclidean).

Figure 1: An illustration of Euclidean vs. non-Euclidean graphs.

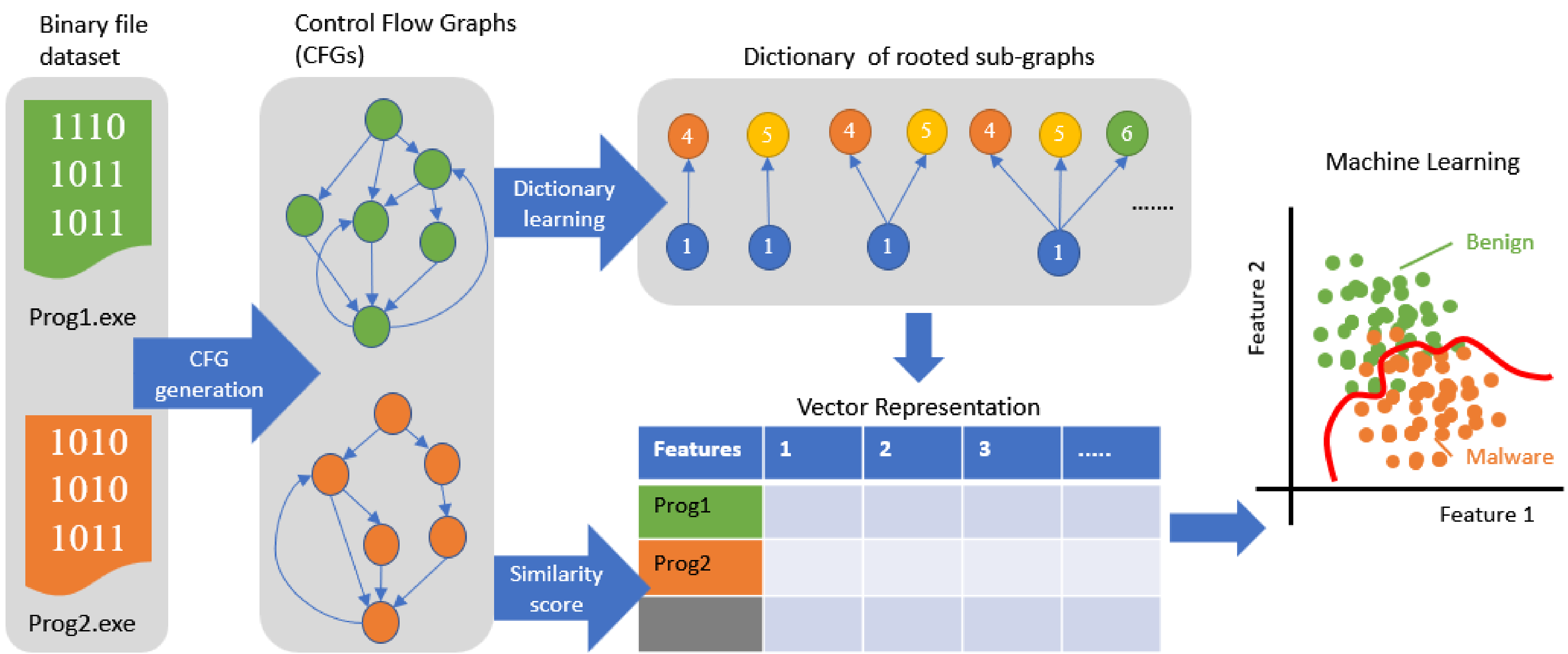


Vector Representation

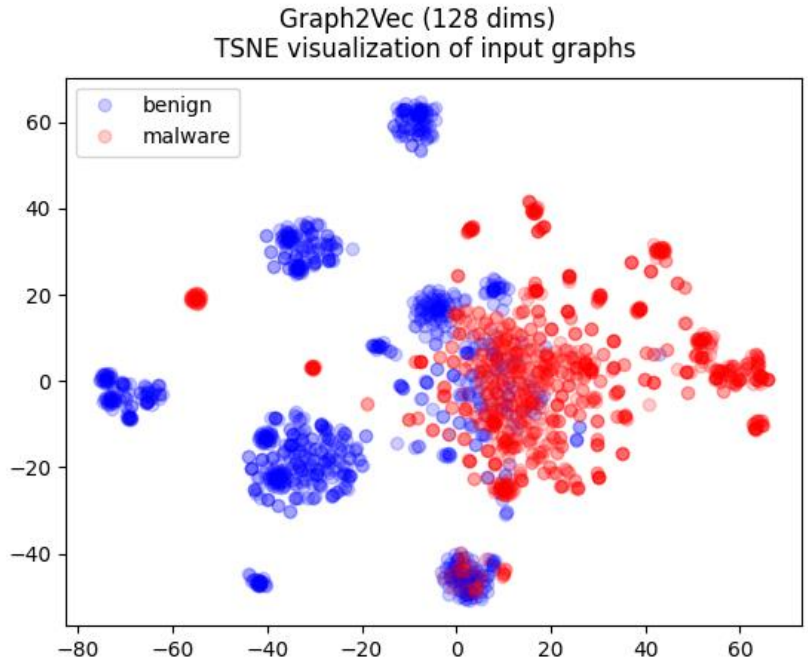
Dimension	Prog1	Prog2
1		
2		
3		
4		
5		
6		
..		

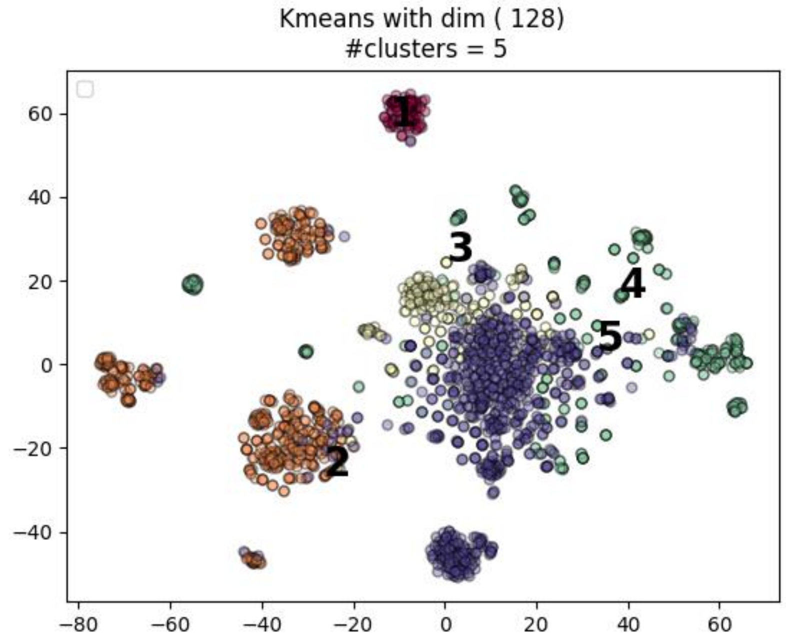
1. Cai, H., Zheng, V.W. and Chang, K.C.C., 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), pp.1616-1637.

Overview of binary file analysis using CFGs



Unsupervised clustering of binary files using CFGs





Goal

Incorporate Sparse representation and its features for whole graph embedding.

A Graph is represented as one sparse vector and two graphs with similar sub- structure are embedded to be closer.

The Graph can be

- Directed
- Cyclic
- Have node feature

Objectives

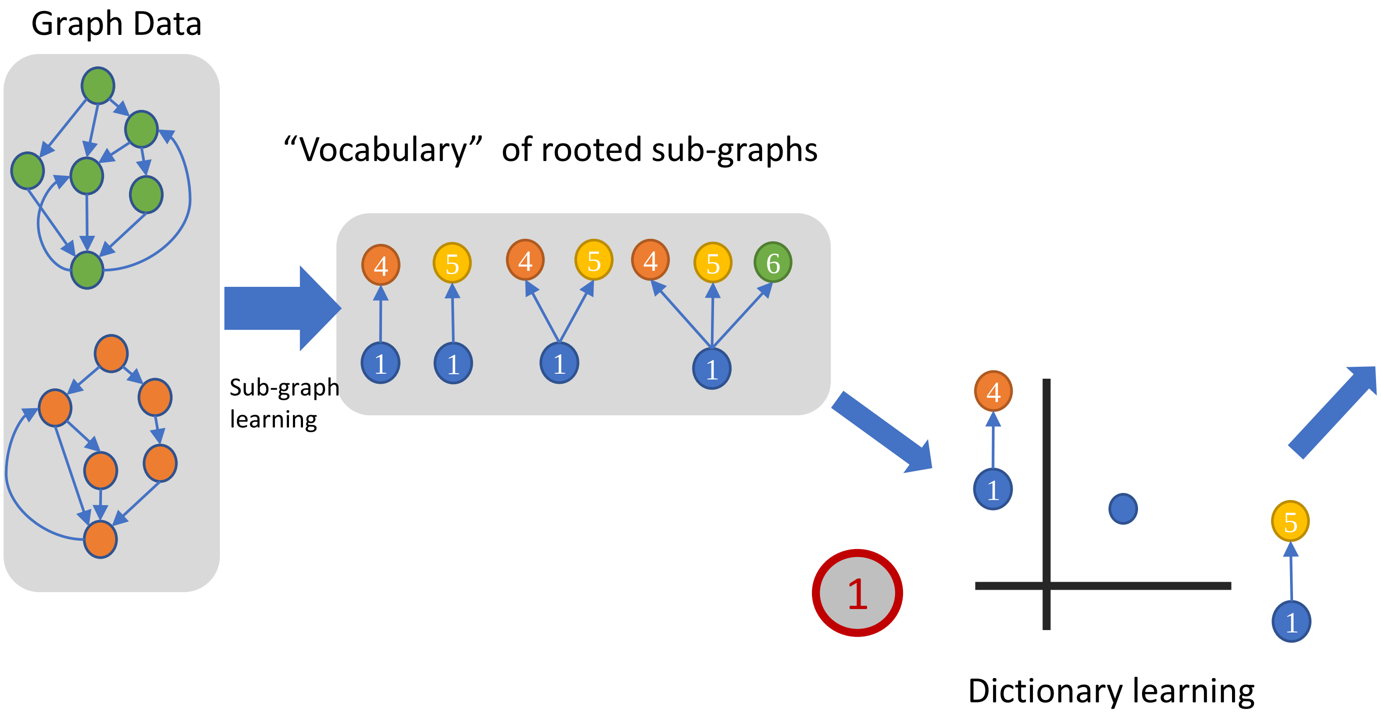
1. To develop a sub-tree pattern based sparse graph embedding method

    WL+KSVD

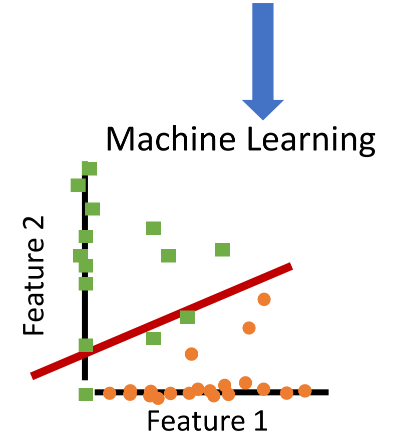
2. Framework for Identifying important sub-tree patterns

3. Develop a sparse graph representation in hyperbolic space

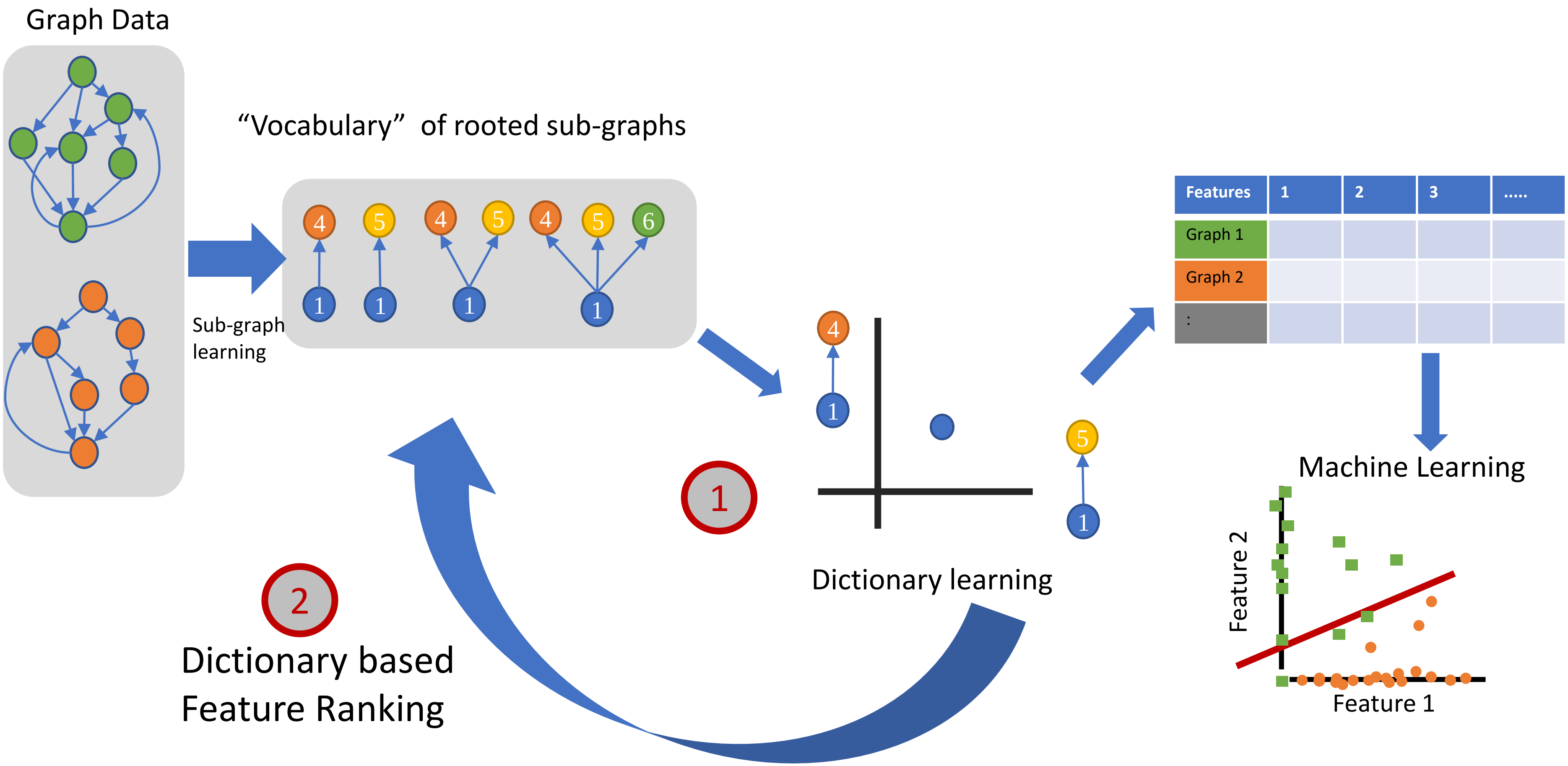
Objectives overview



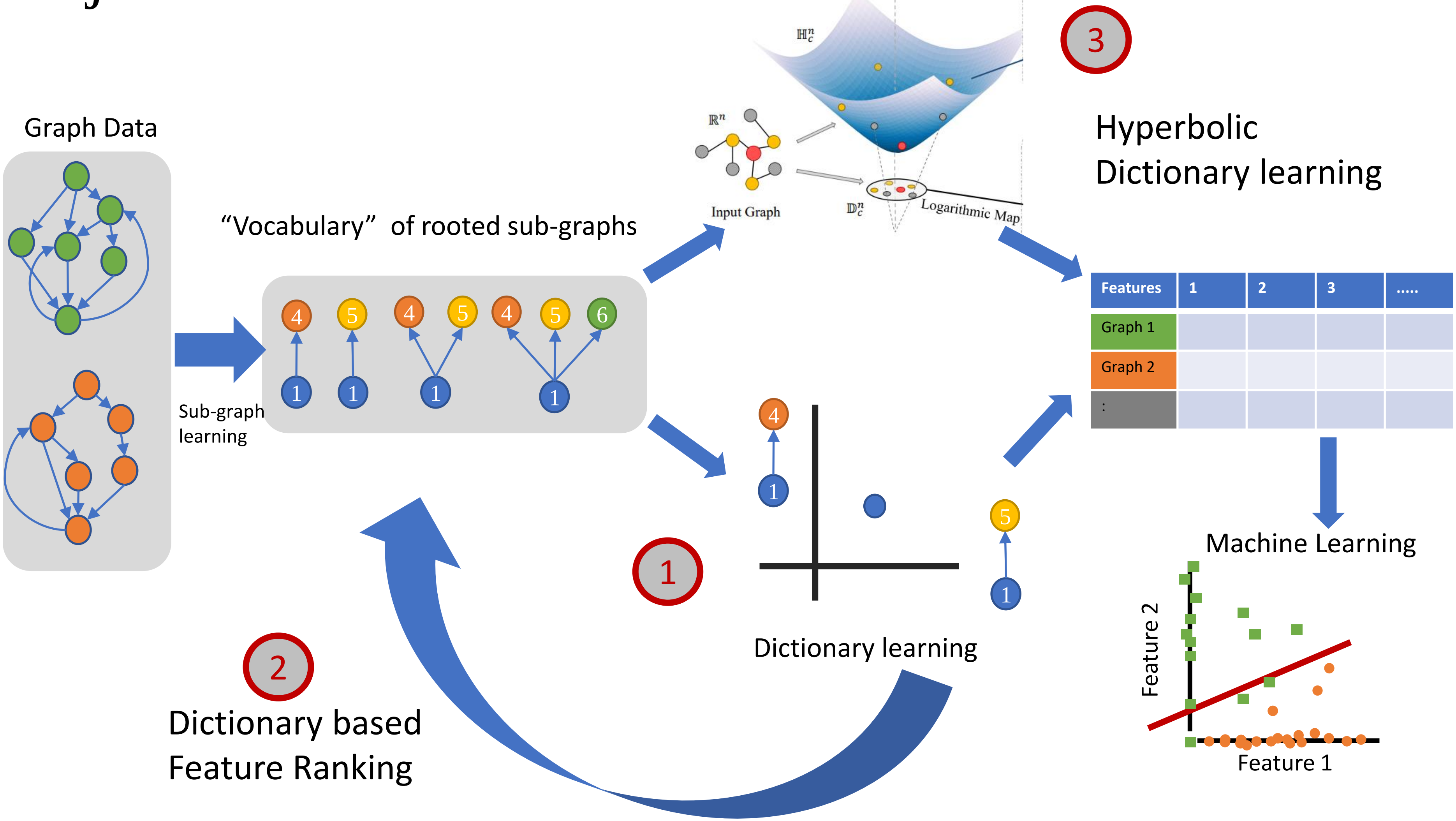
Features	1	2	3	.....
Graph 1				
Graph 2				
:				



Objectives overview



Objectives overview



Overview

- Introduction
  - Motivation / Problem
  - Goal
  - Objectives
- Background
  - Graph embedding
  - Sparse representation
  - Feature ranking
- Workflow
  - Preliminary work
  - Proposed timeline
  - Future directions

Graph embedding

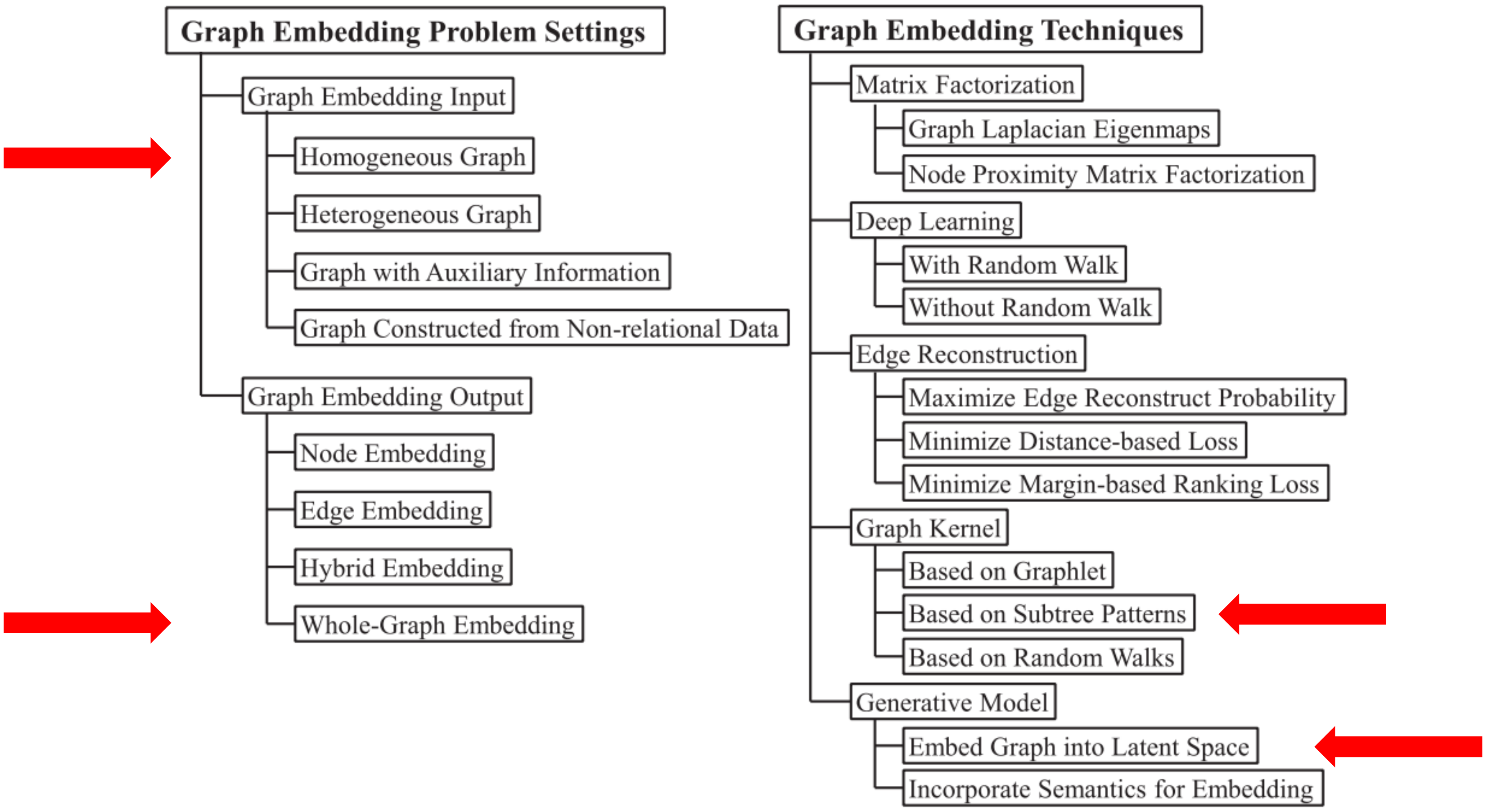
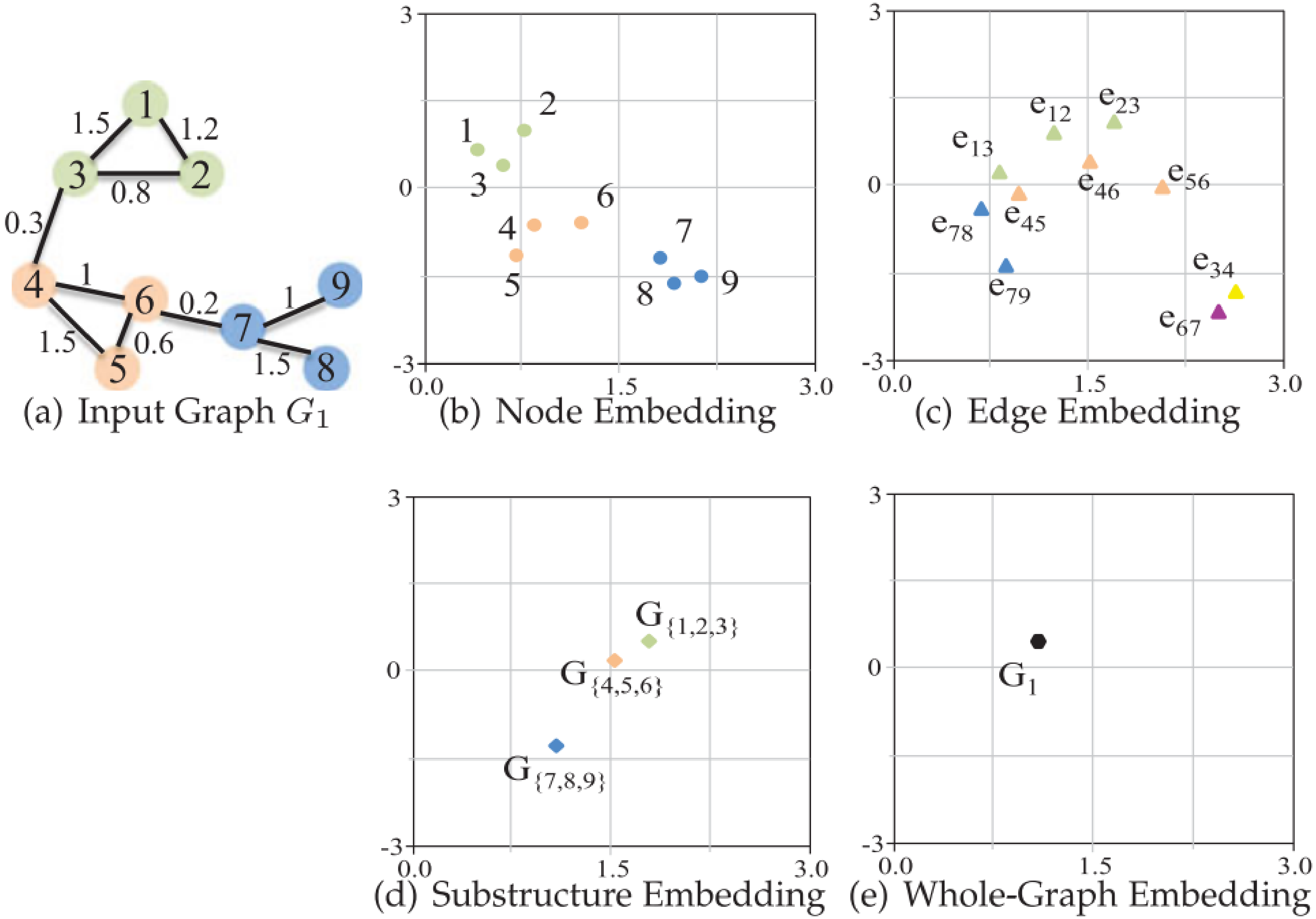


Fig. 2. Graph embedding taxonomies by problems and techniques.

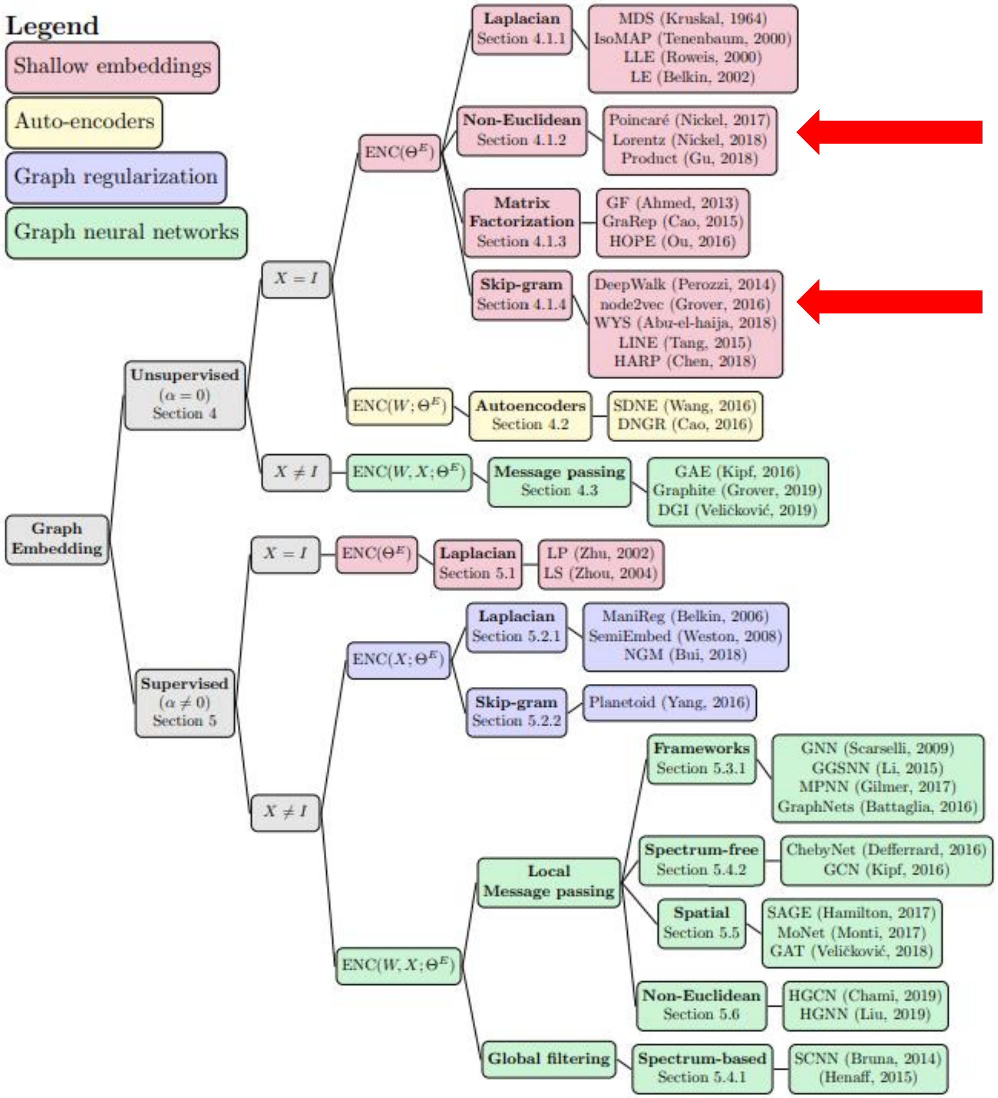
[1] Cai, H., Zheng, V.W. and Chang, K.C.C., 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), pp.1616-1637.

Graph embedding types



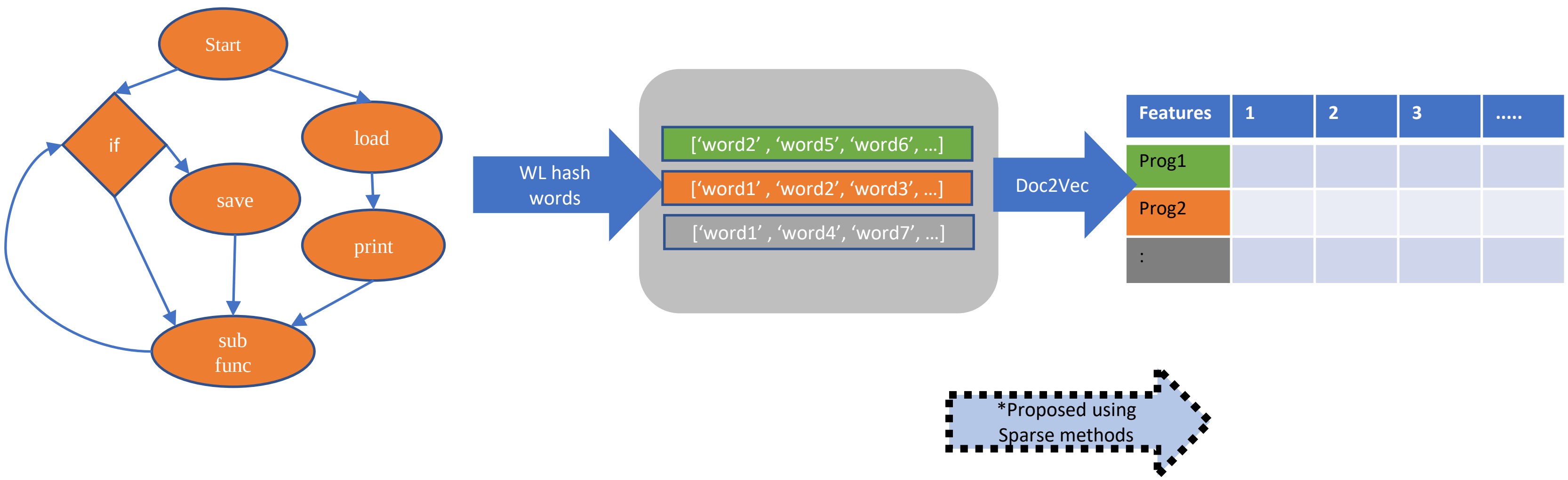
[1] Cai, H., Zheng, V.W. and Chang, K.C.C., 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), pp.1616-1637.

Graph embedding techniques

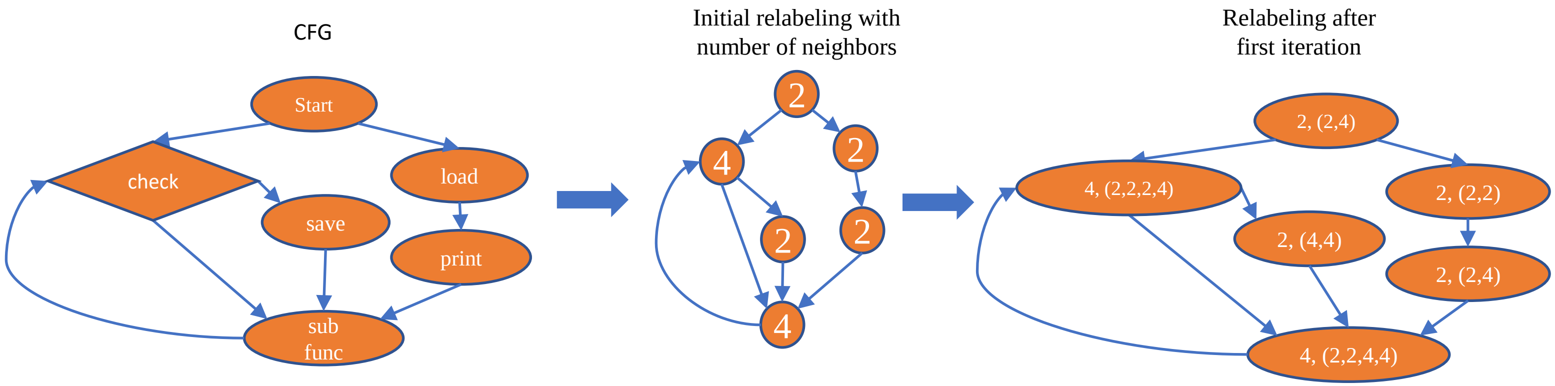


[1] Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C. and Murphy, K., 2022. Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, *23*(89), pp.1-64.

Graph2Vec overview

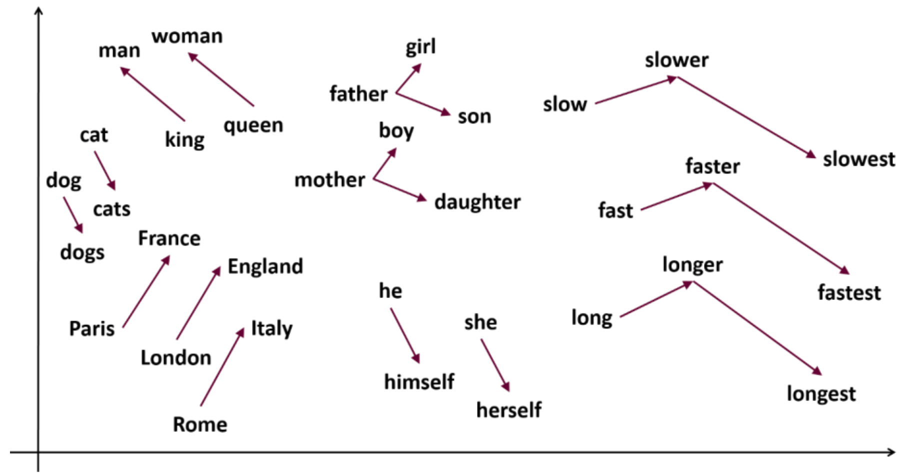


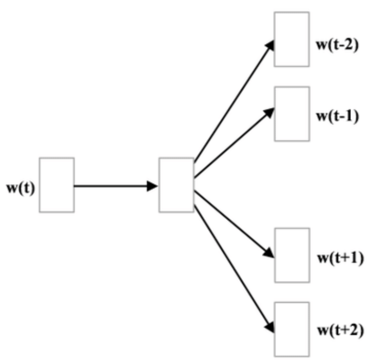
Weisfeiler-Lehman subtree hash

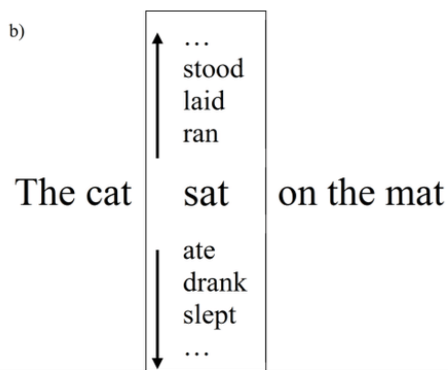


Node	Word
Start	2, hash(2,4)
check	4, hash(2,2,2,4)
Load	2, hash (2,2)
Save	2, hash(4,4)
Print	2, hash(2,4)
Sub_func	4, hash(2,2,4,4)
…	….

Word2Vec







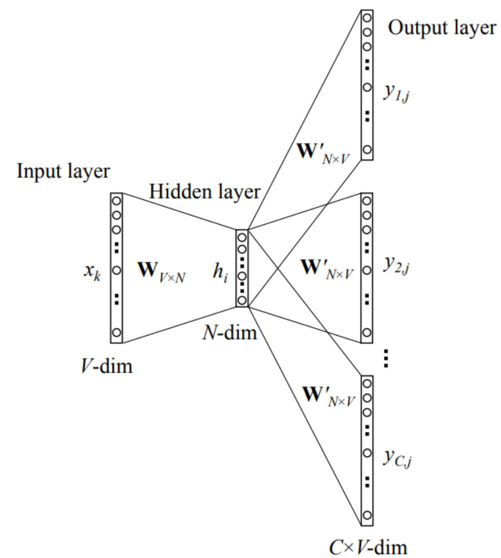
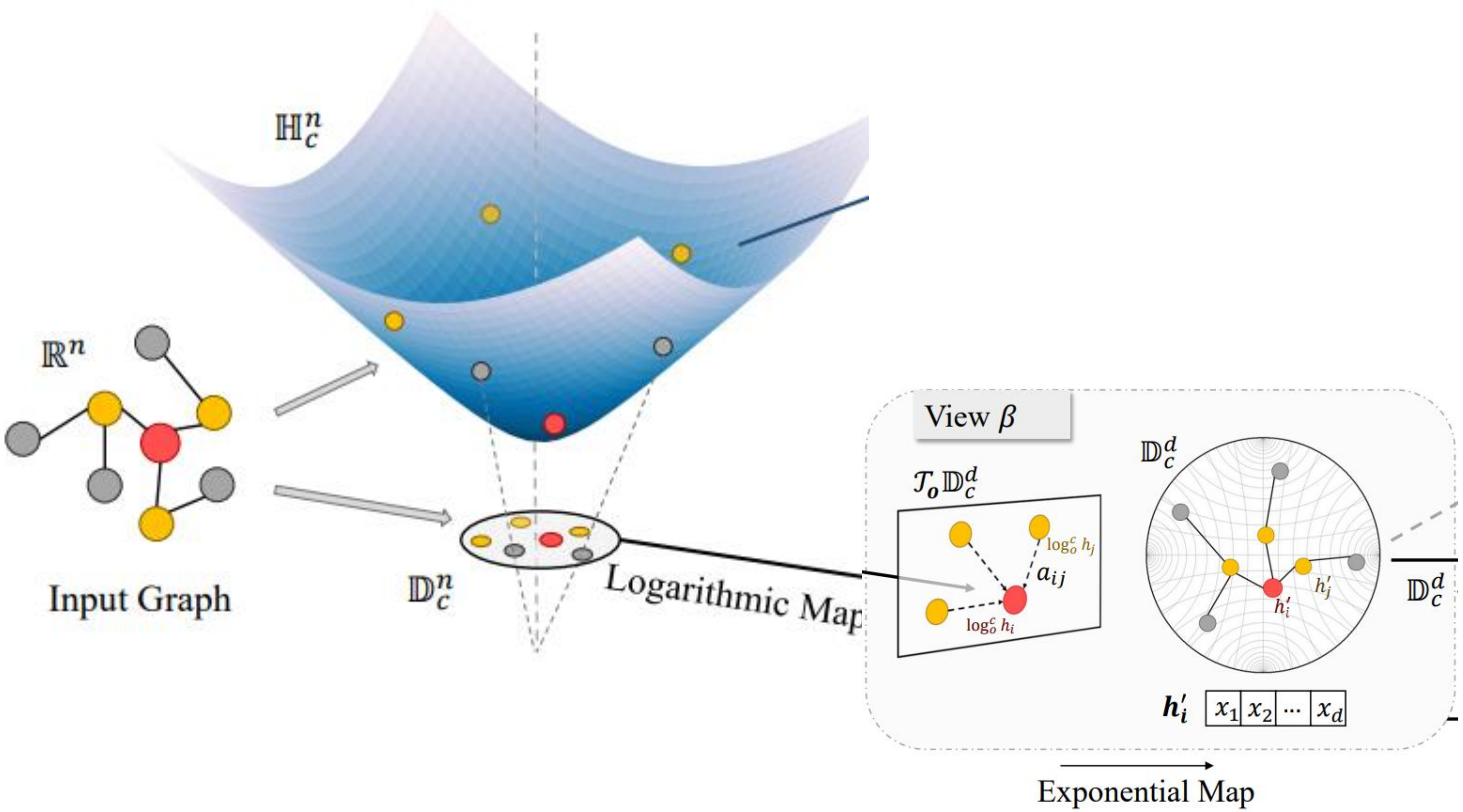


Figure 3: The skip-gram model.

https://ronxin.github.io/wevi/

1. Rong, X., 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

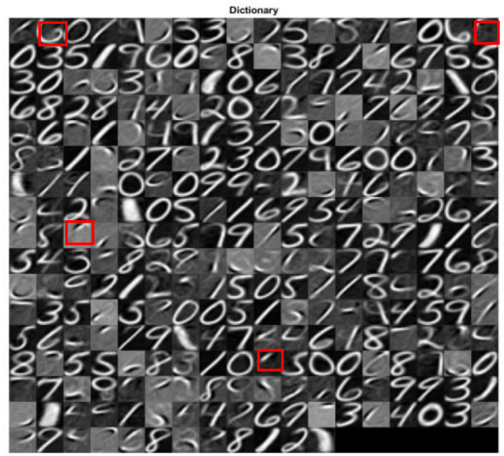
Poincare map



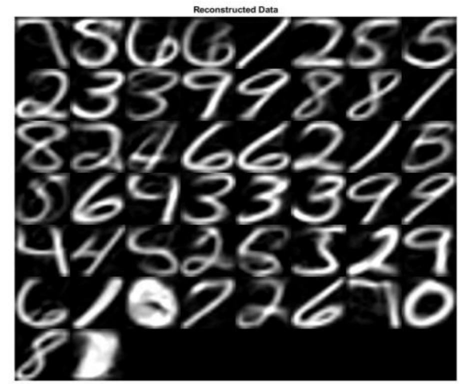
1. Liu, J., Yang, M., Zhou, M., Feng, S. and Fournier-Viger, P., 2022. Enhancing hyperbolic graph embeddings via contrastive learning. *arXiv preprint arXiv:2201.08554*.

Sparse Representation

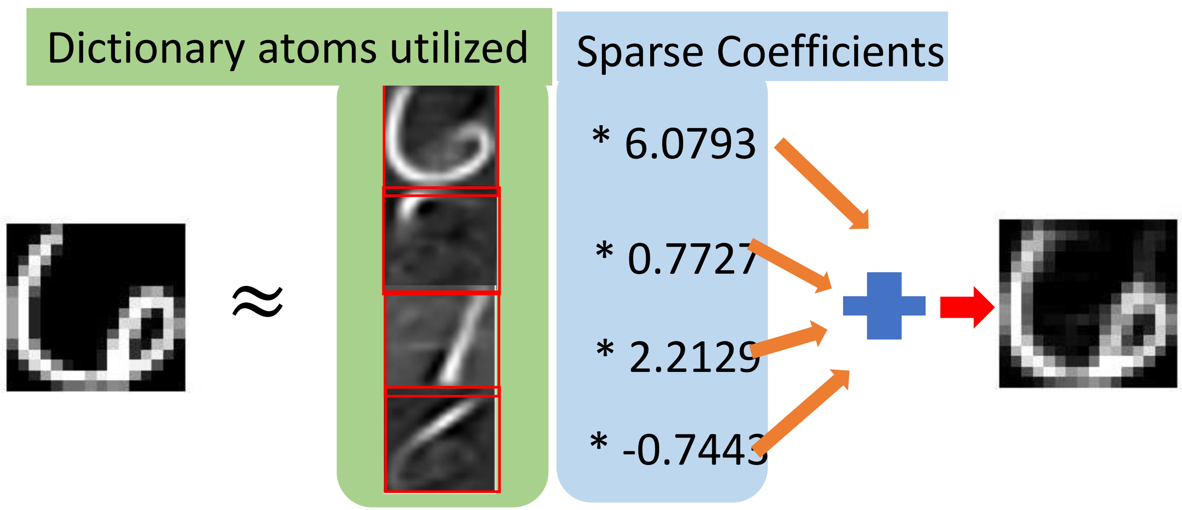
Input Data



Learned dictionary



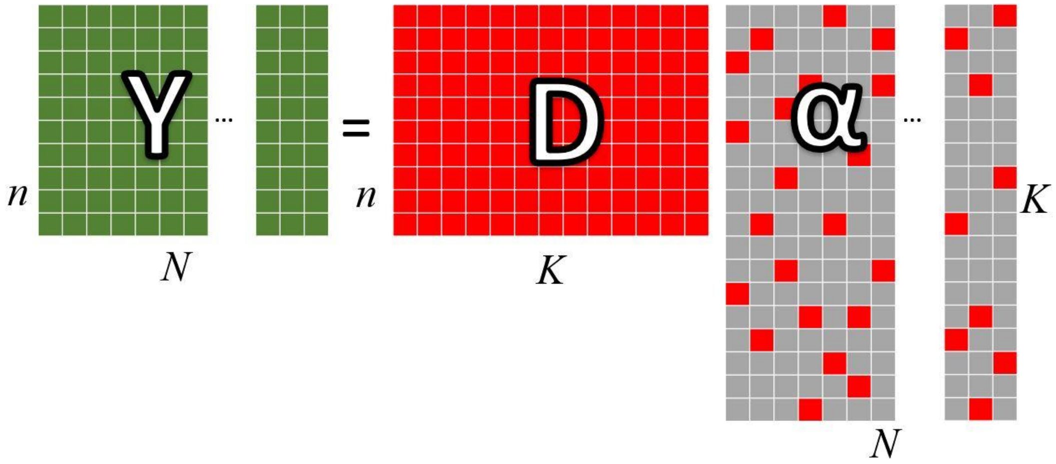
Reconstructed Data



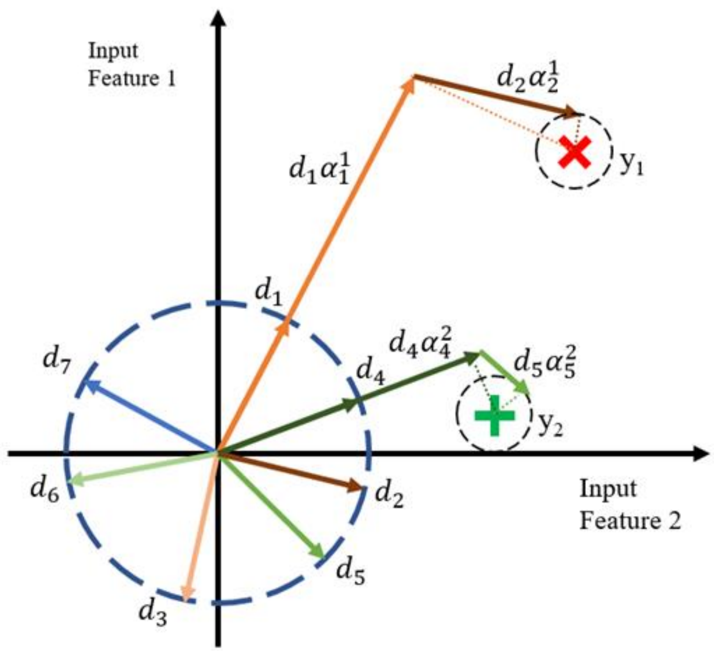
Sparse Representation

- Input signals : $\boldsymbol{Y} = [y_1, y_2, \dots, y_N] \in R^{n \times N}$
- Dictionary elements : $d_i \in R^n$
- Dictionary : $\boldsymbol{D} = [d_1, d_2, \dots, d_K] \in R^{n \times K}$
- Resulting signal : $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_N] \in R^{K \times N}$
- Sparsity : S

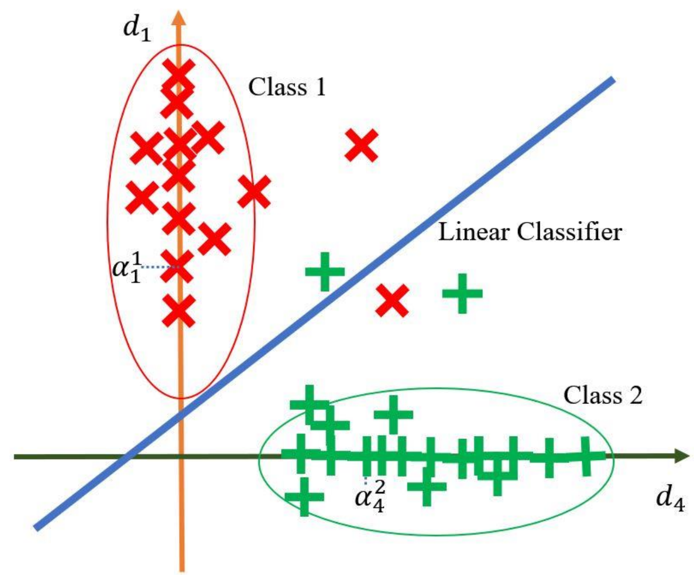
$$\boxed{\underset{D,\alpha}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{D}\alpha||_2^2 \;\; s.t. \forall i, ||\alpha_i||_0 \leq S,}$$



Learned dictionary is in the same input space



Input
space



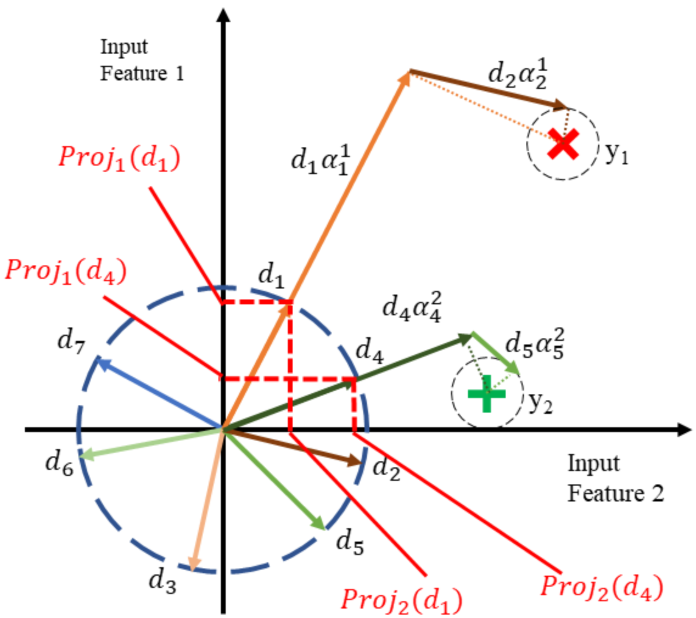
Dictionary
space

1. Liyanage, K. and Whitaker, B.M., 2022, May. Feature Analysis in Satellite Image Classification Using LC-KSVD and Frozen Dictionary Learning. In *2022 Intermountain Engineering, Technology and Computing (IETC)* (pp. 1-6). IEEE.

Dictionary based Feature Ranking metrics

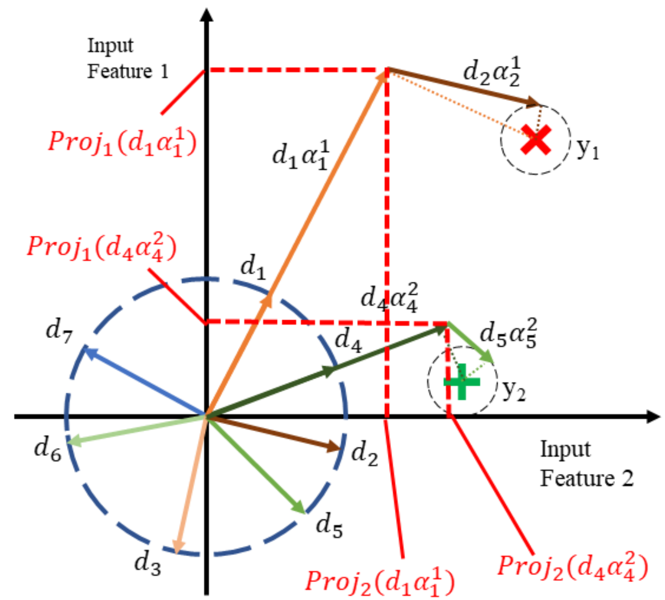
Dictionary mapping

$$\mathbf{D}_{\text{map}}(j) = \sum_{i=1}^{m} \text{Proj}_j(D_i)^2 = \sum_{i=1}^{m} \mathbf{D}^2_{(j,i)}$$



Dictionary utilization

$$\mathbf{D}_{\text{util}}(j) = \sum_{i=1}^{m} \text{Proj}_j\left(D_i \cdot \sum_{k=1}^{n} |\alpha_{j,k}|\right)$$



Sparse coding-based FR

Pros

Model agnostic

- Calculated Sparse coefficients can be used with different models

Simple relationship

- More intuitive mapping

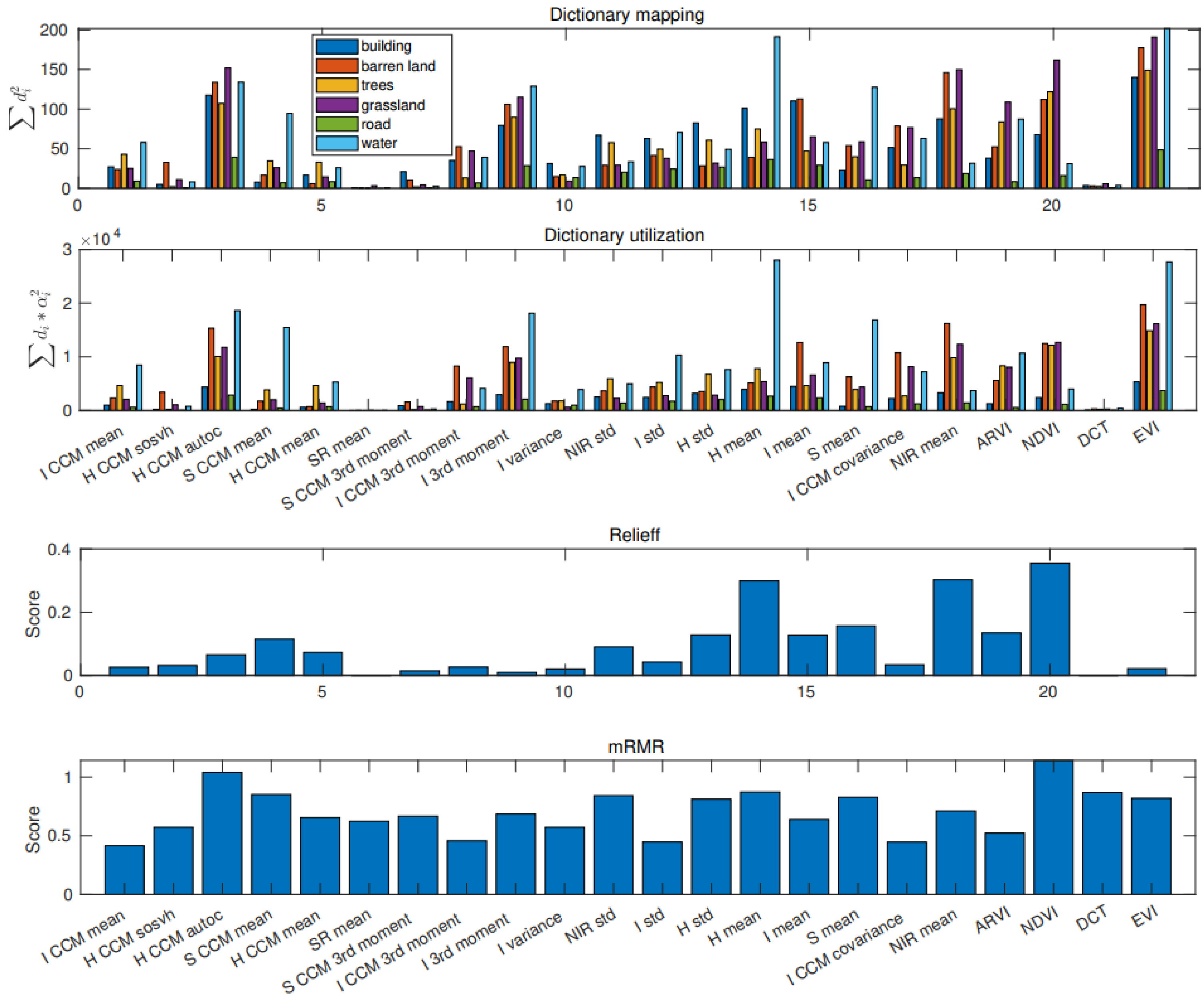
Non-myopic manner

- Take into consideration correlations and redundancies

Cons

Time must be spent to compute the sparse coefficients

- Time spent learning sparse coefficients is not wasted

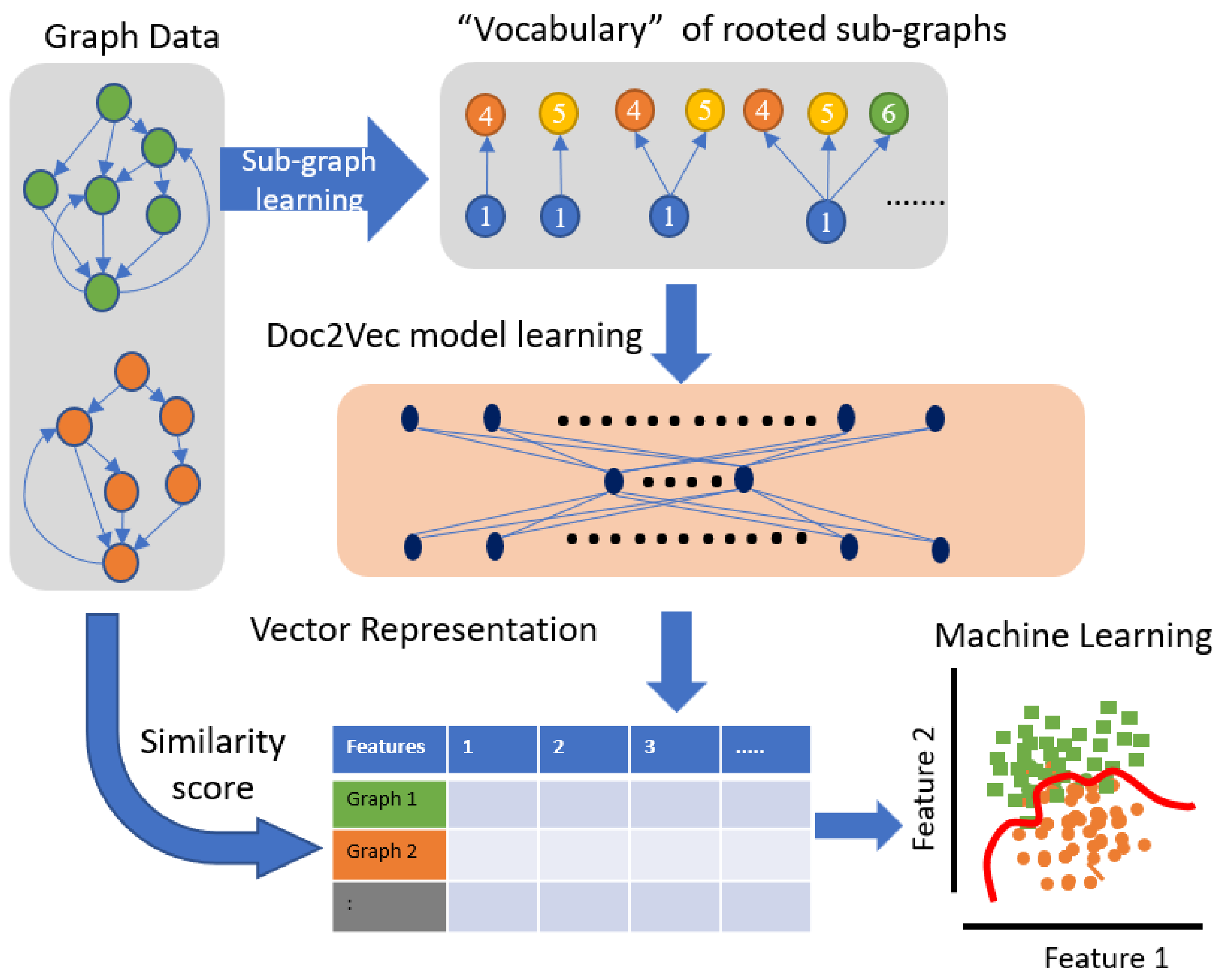


Dictionary mapping
building
barren land
trees
grassland
road
water
Dictionary utilization
I CCM mean
H CCM sosvh
H CCM autoc
S CCM mean
H CCM mean
SR mean
S CCM 3rd moment
I CCM 3rd moment
I 3rd moment
I variance
NIR std
I std
H std
H mean
I mean
S mean
I CCM covariance
NIR mean
ARVI
NDVI
DCT
EVI
Relieff
Score
mRMR
Score

Overview

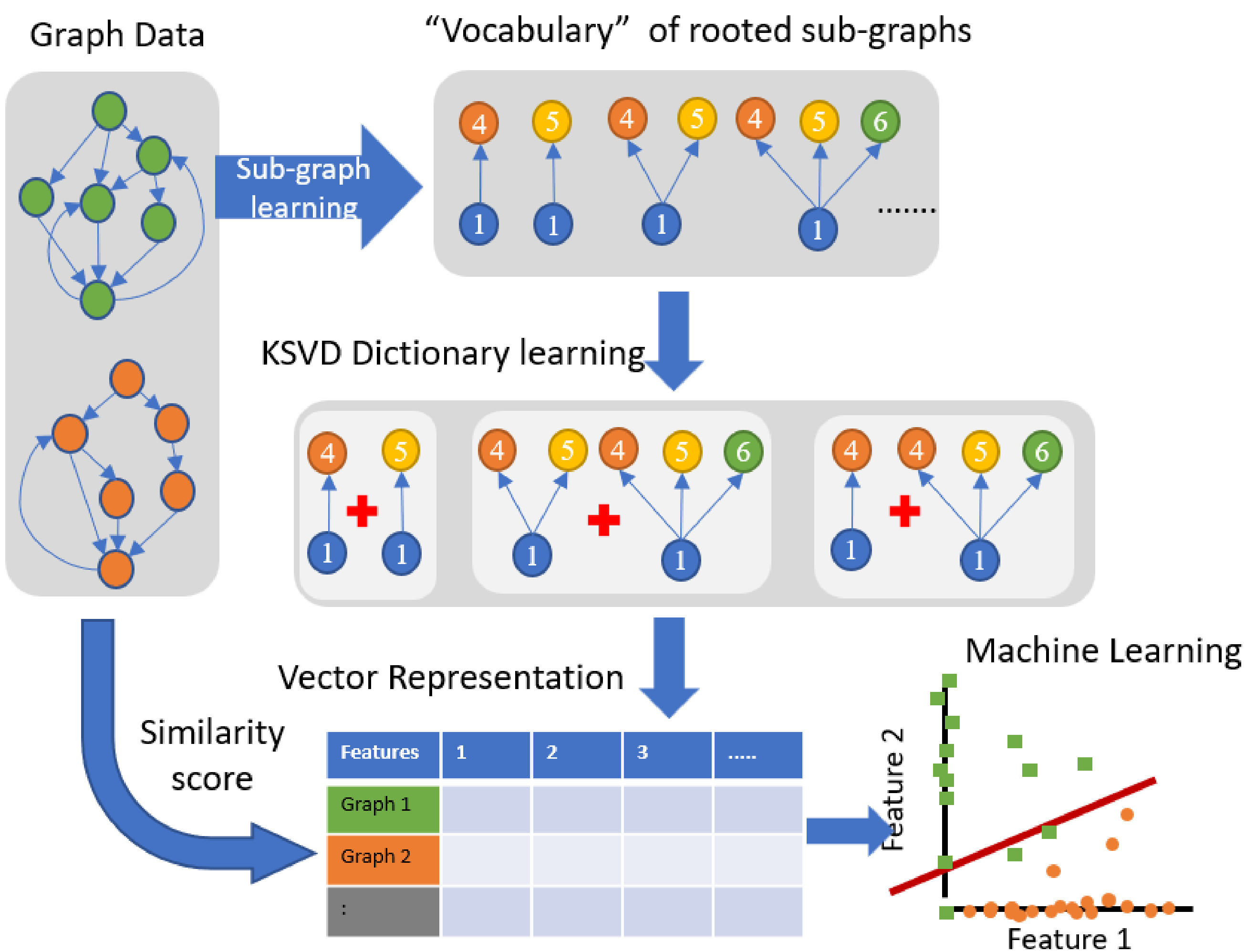
- Introduction
  - Motivation / Problem
  - Goal
  - Objectives
- Background
  - Graph embedding
  - Sparse representation
  - Feature ranking
- Workflow
  - Preliminary work
  - Proposed timeline
  - Future directions

Graph2Vec



Graph Data
“Vocabulary” of rooted sub-graphs
Sub-graph learning
4
5
4
5
4
5
6
1
1
1
1
....
Doc2Vec model learning
Vector Representation
Similarity score
Features
1
2
3
.....
Graph 1
Graph 2
:
Machine Learning
Feature 2
Feature 1

WL+KSVD



Graph Data
"Vocabulary" of rooted sub-graphs
Sub-graph learning
KSVD Dictionary learning
Vector Representation
Similarity score
Features
1
2
3
.....
Graph 1
Graph 2
:
Machine Learning
Feature 2
Feature 1

Preliminary results

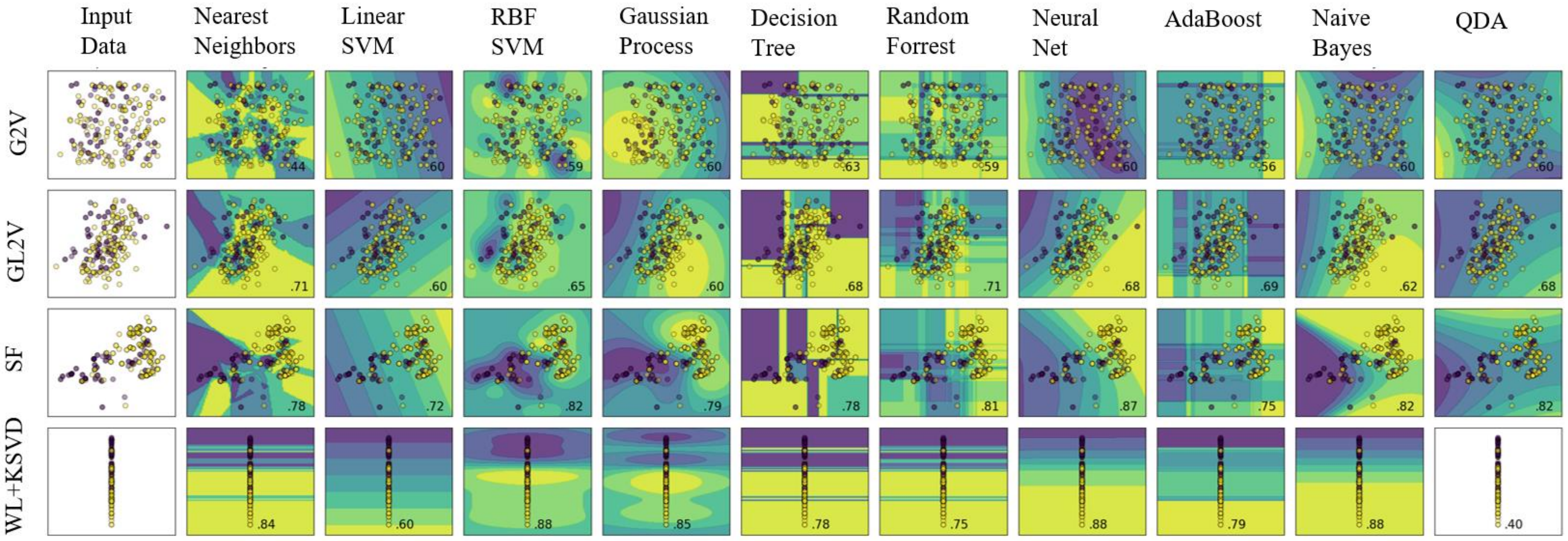
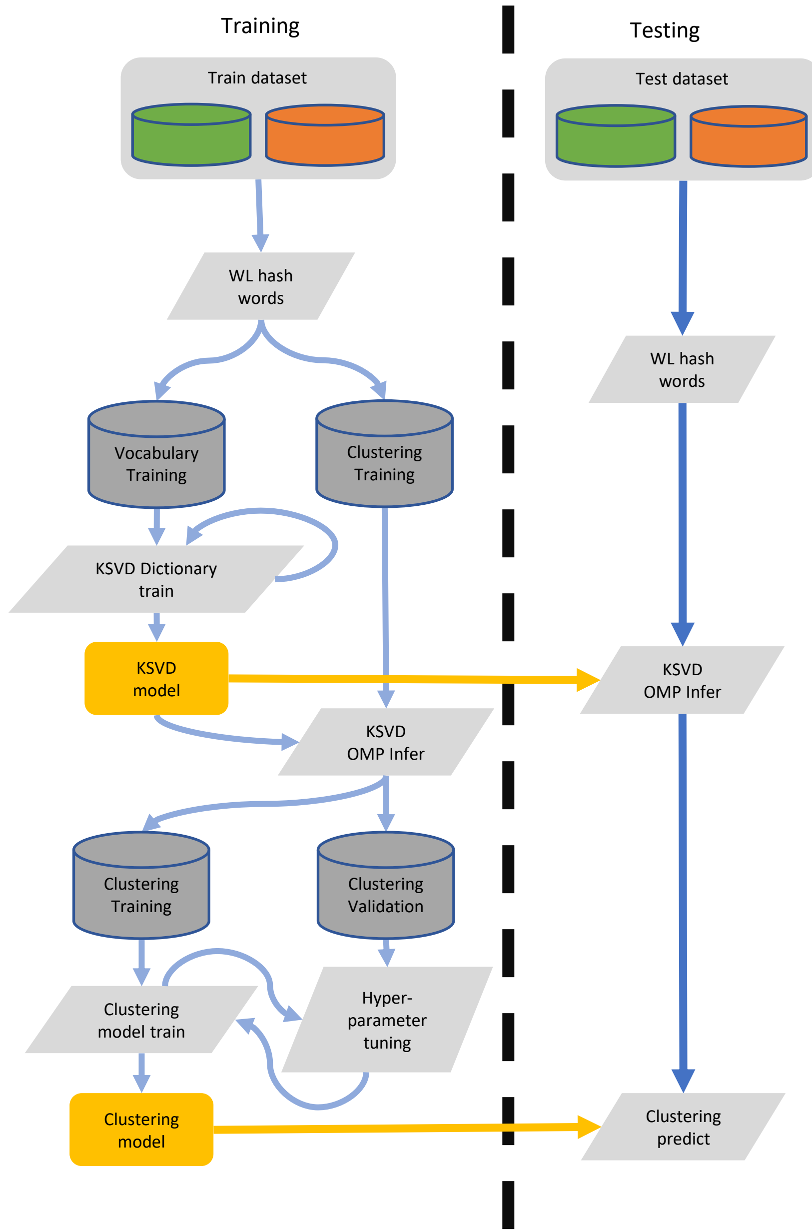


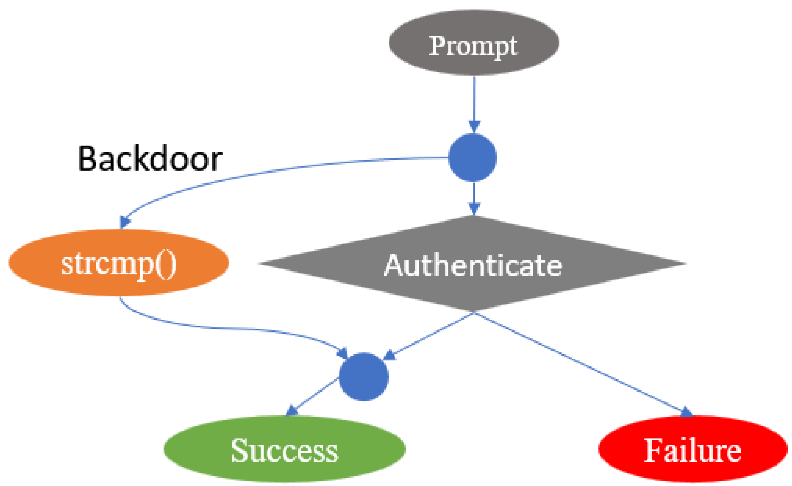
Table 2. Linear SVM accuracy with $N = 1024$ embedding

	MU	PTC	PROT	NCI1	NCI109
G2V	68.55 $\pm$10.03	55.23 $\pm$5.9	67.30 $\pm$0.87	59.30 $\pm$4.46	56.46 $\pm$2.82
GL2V	74.92 $\pm$7.8	52.04 $\pm$6.5	69.09 $\pm$1.38	64.52 $\pm$1.99	62.98 $\pm$2.97
SF	83.47 $\pm$4.15	57.59 $\pm$9.34	70.98 $\pm$1.00	61.90 $\pm$3.24	61.96 $\pm$2.40
WL+ KSVD	72.38 $\pm$3.20	54.36 $\pm$2.31	64.60 $\pm$2.00	64.16 $\pm$2.19	62.93 $\pm$0.24

Pipeline


Training
Train dataset
WL hash words
Vocabulary Training
Clustering Training
KSVD Dictionary train
KSVD model
KSVD OMP Infer
Clustering Training
Clustering Validation
Clustering model train
Hyper-parameter tuning
Clustering model
Testing
Test dataset
WL hash words
KSVD OMP Infer
Clustering predict

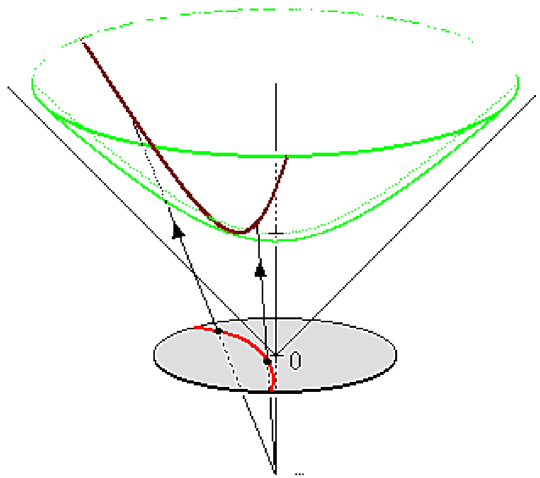
Identifying vulnerable subtree structures

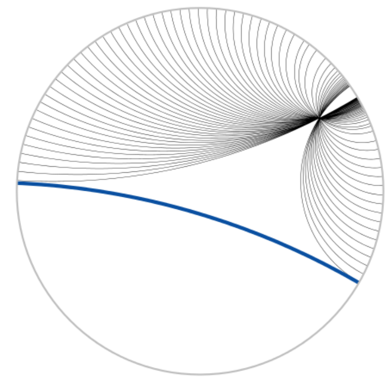


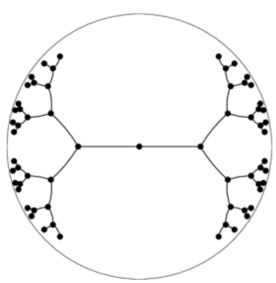
Conceptual authentication bypass vulnerability

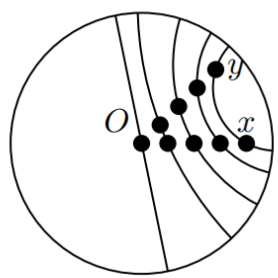
1. https://www.ndss-symposium.org/wp-content/uploads/2017/09/11Firmalice.slide_.pdf

Hyperbolic space









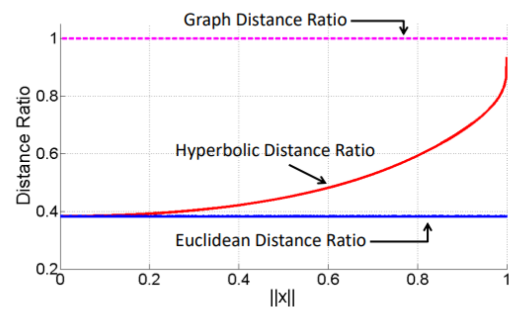


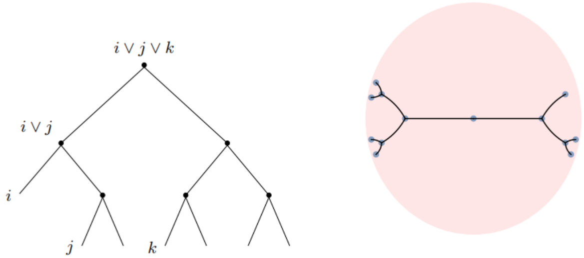
Figure 1. Left: Embedding of a binary tree in the Poincaré disk. Right: Geodesics and distances. As x and y move towards the outside of the disk (i.e., letting $\|x\|, \|y\| \to 1$), the distance $d_H(x, y)$ approaches $d_H(x, O) + d_H(O, y)$.

1. Sala, F., De Sa, C., Gu, A. and Ré, C., 2018, July. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning* (pp. 4460-4469). PMLR.

Hyperbolic space dictionary learning

Subtree embedding in Hyperbolic

Chami, I., Gu, A., Chatziafratis, V. and Ré, C., 2020. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, *33*, pp.15065-15076.



(a) (b)

SVD in hyperbolic

Onn, R., Steinhardt, A.O. and Bojanczyk, A., 1989, August. The hyperbolic singular value decomposition and applications. In *Proceedings of the 32nd Midwest Symposium on Circuits and Systems,* (pp. 575-577). IEEE.

K-means in hyperbolic

Djeddal, H., Touzari, L., Giovanidis, A., Phung, C.D. and Secci, S., 2021. Hyperbolic K-means for traffic-aware clustering in cloud and virtualized RANs. *Computer Communications*, *176*, pp.258-271.

Pursuit algorithms in hyperbbolic

Tabaghi, P. and Dokmanić, I., 2020, August. Hyperbolic distance matrices. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1728-1738).

Sonthalia, R. and Gilbert, A., 2020. Tree! i am no tree! i am a low dimensional hyperbolic embedding. *Advances in Neural Information Processing Systems*, *33*, pp.845-856.

Q&A
Summary

Objectives

Sparse dictionary learning on subtree patterns
Feature ranking of subtree patterns
Sparse representation on Hyperbolic space

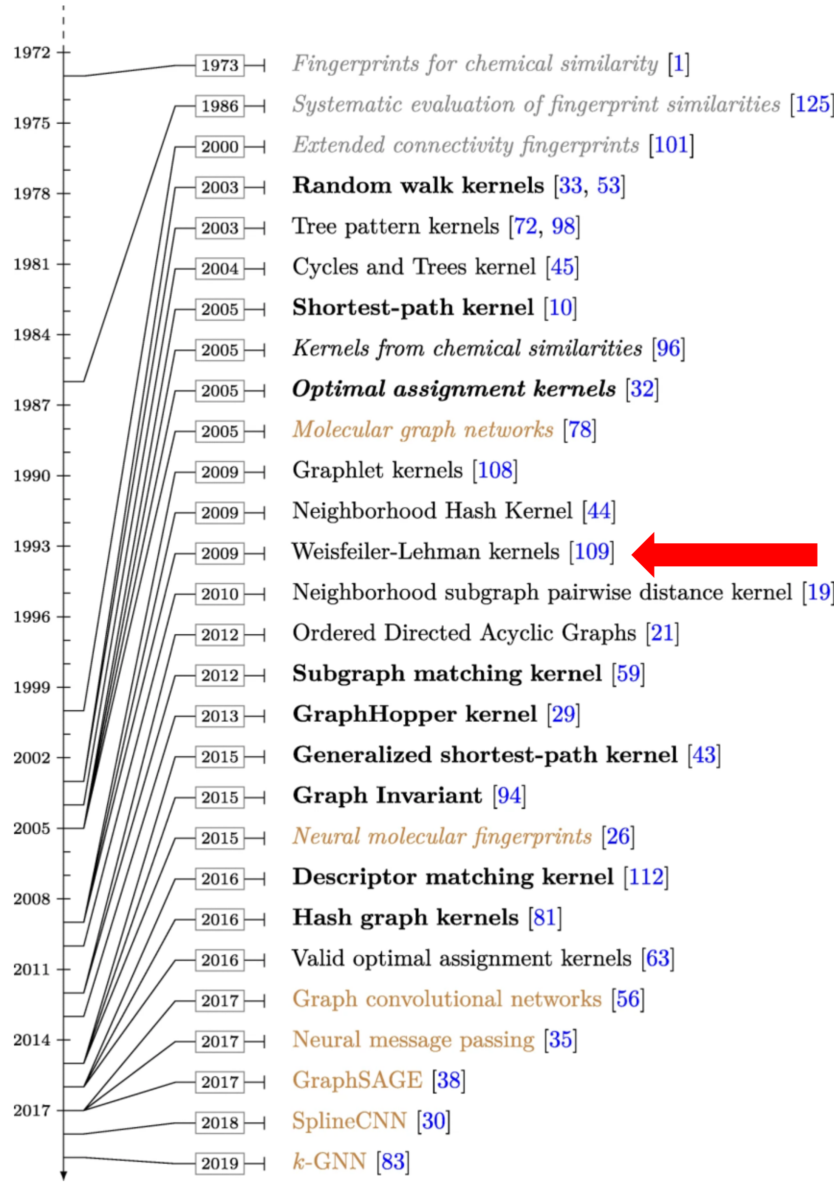
Advantage

Linear relationship
Simpler (Low order) ML model
Intuitive

Outcomes

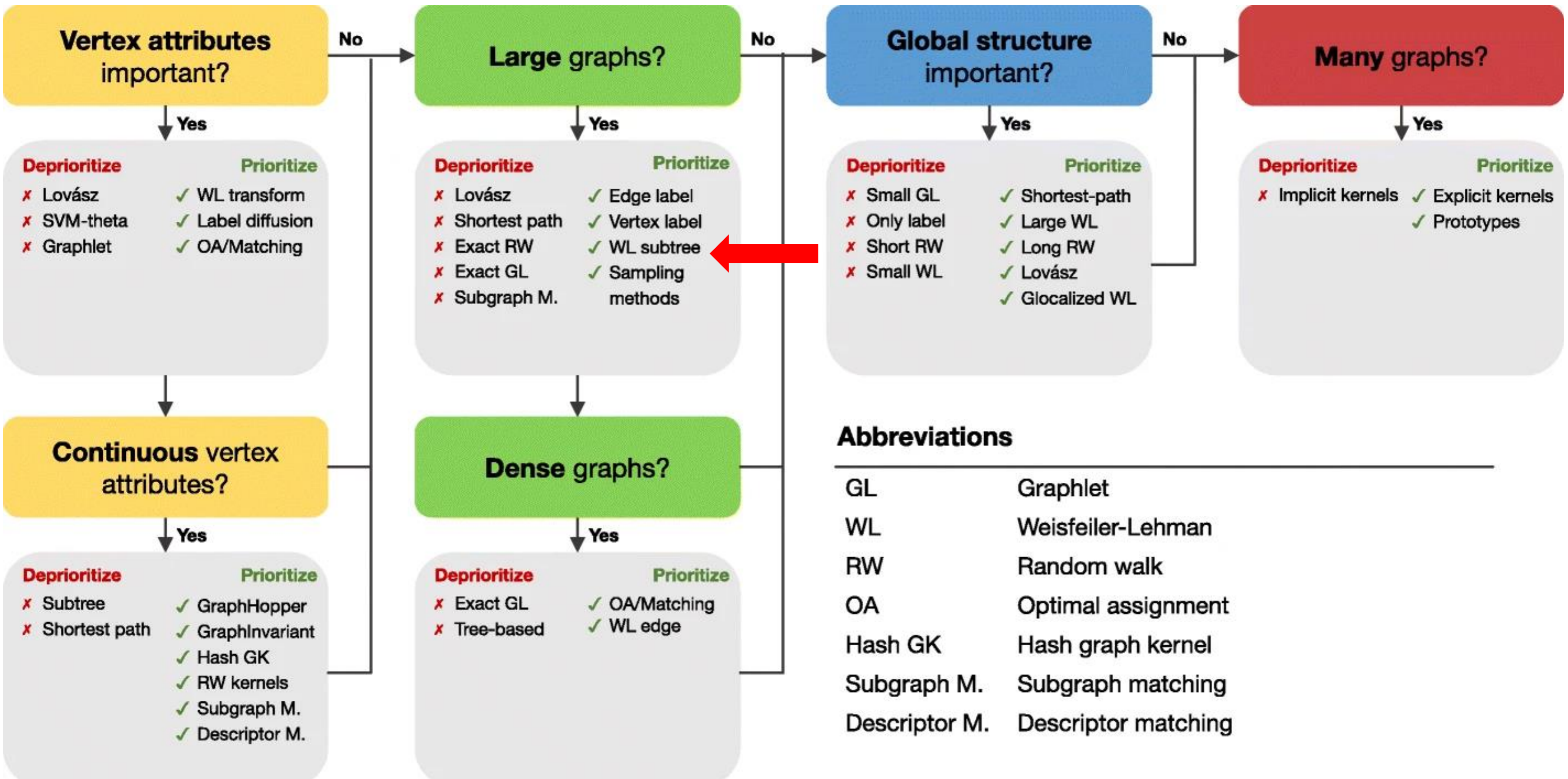
Publicly available python package

Graph Kernels



1. Kriege, N.M., Johansson, F.D. and Morris, C., 2020. A survey on graph kernels. Applied Network Science, 5(1), pp.1-42.

Graph Kernel selection guideline

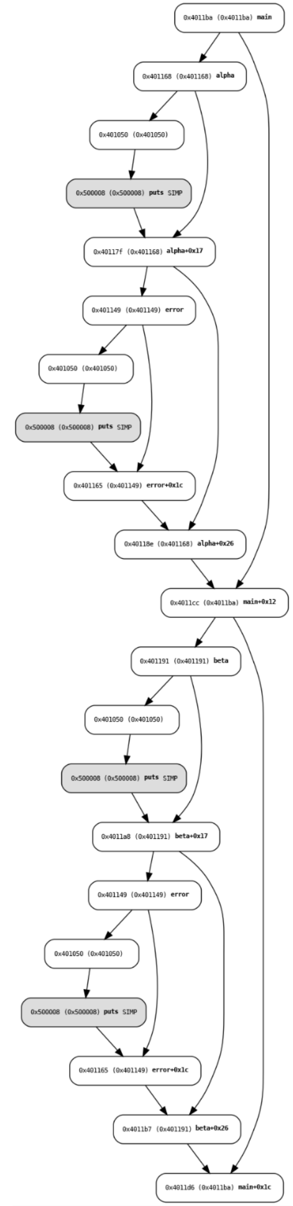


Guidelines for prioritizing kernels for consideration based on known properties of the graph learning problem.

1. Kriege, N.M., Johansson, F.D. and Morris, C., 2020. A survey on graph kernels. Applied Network Science, 5(1), pp.1-42.

Typical CFG for a given program

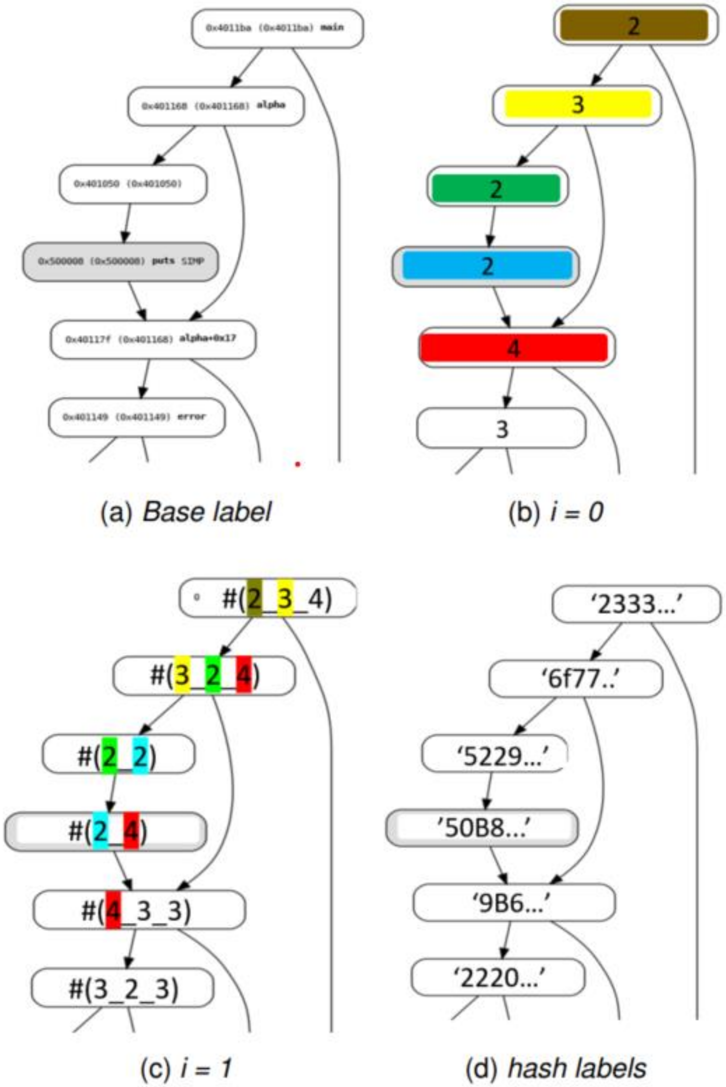
```
void error(char *error)
{
   puts(error);
}

void alpha()
{
  puts("alpha");
  error("alpha!");
}

void beta()
{
  puts("beta");
  error("beta!");
}

void main()
{
  alpha();
  beta();
}
```



Conceptual authentication bypass vulnerability

1. https://www.ndss-symposium.org/wp-content/uploads/2017/09/11Firmalice.slide_.pdf

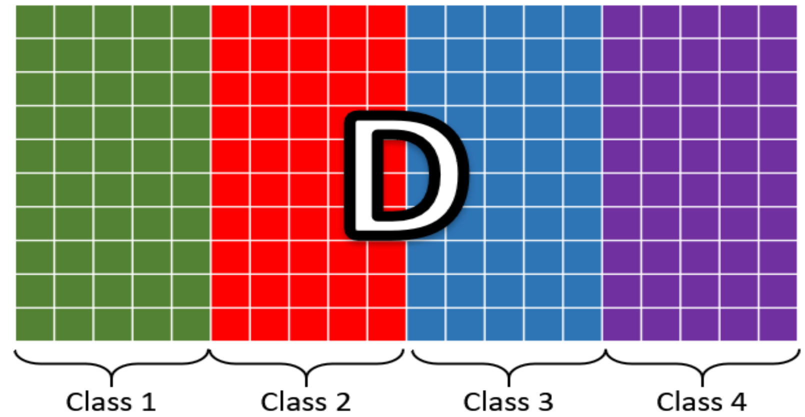
WL subtree hash relabeling



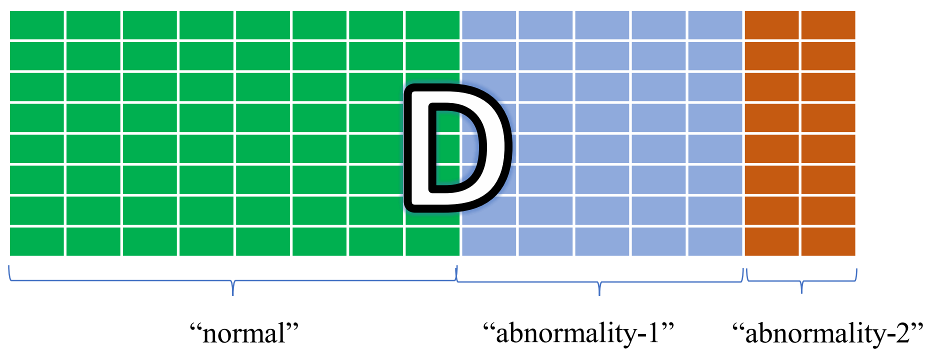
WL graph hash algorithms' first iteration in the context of CFG

Supervised dictionary learning

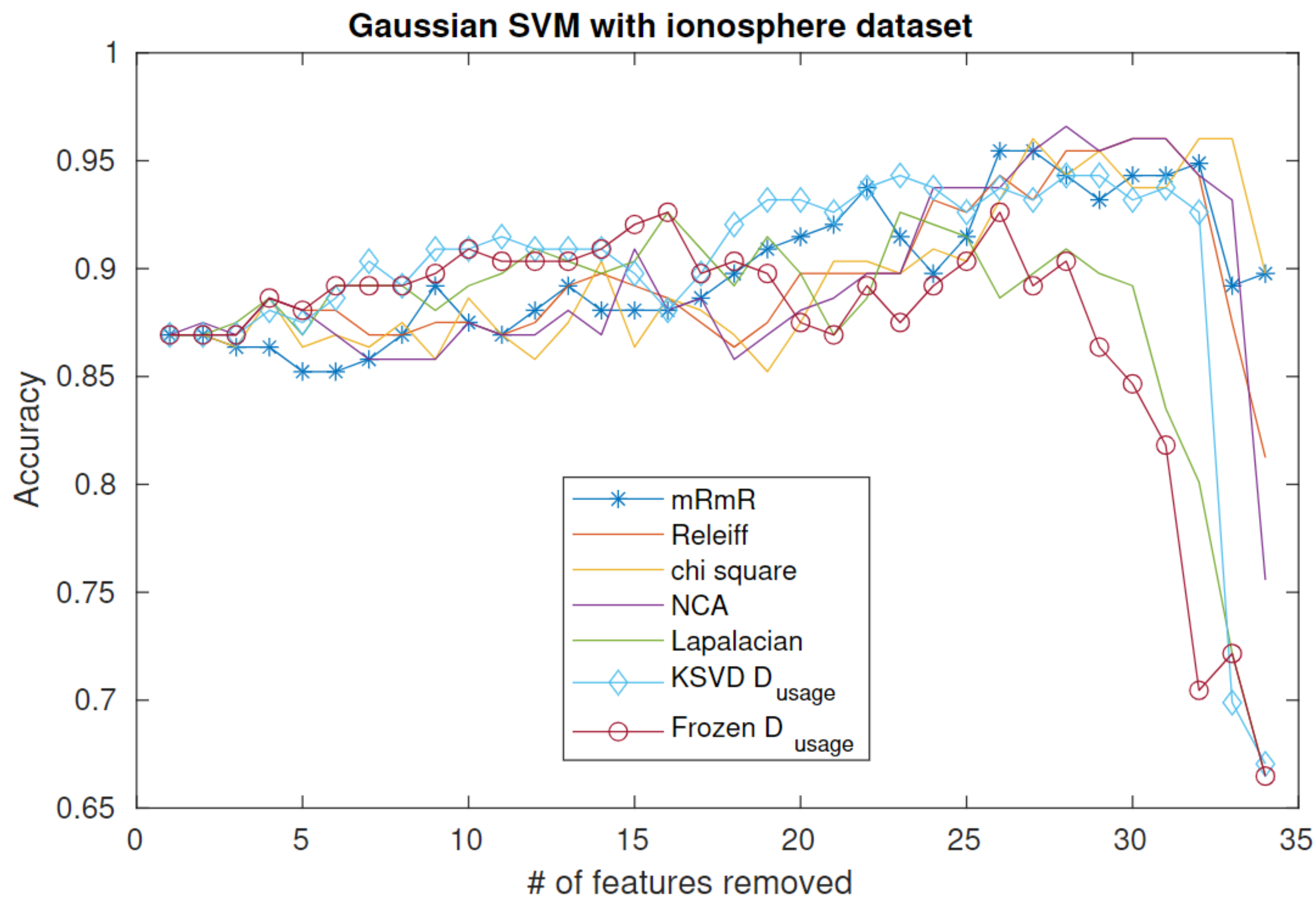
Label Consistent-KSVD



Frozen-KSVD



Feature selection



Overview

Introduction

Motivation

Problem

Goal

Background

Graph embedding

Sparse representation

Feature ranking

Objectives

Sparse dictionary learning on subtree patterns

Feature ranking of subtree patterns

Sparse representation on Hyperbolic space