Introduction

Welcome to this lab! At this lab, we will learn:

1. How to build a graph from a file or create a simple graph by ourself
2. Implement DeepWalk in the simplest way based on the paper DeepWalk.

Exercise

Download data and install packages

In [ ]:

```
!gdown --id "1vqsjGzGZnpCEgHliEsVmAvzm9_1h3L-Y&export=download"
!unrar x -Y "/content/lab1.rar" -d "/content/"
```

```
Downloading...
From: https://drive.google.com/uc?id=1vqsjGzGZnpCEgHliEsVmAvzm9_1h3L-Y&export=
download
To: /content/lab1.rar
100% 50.0k/50.0k [00:00<00:00, 30.9MB/s]

UNRAR 5.50 freeware      Copyright (c) 1993-2017 Alexander Roshal


Extracting from /content/lab1.rar

Extracting  /content/lab1_big_edgelist.txt                              99%
OK
Extracting  /content/lab1_small_edgelist.txt                            99%
OK
All OK
```

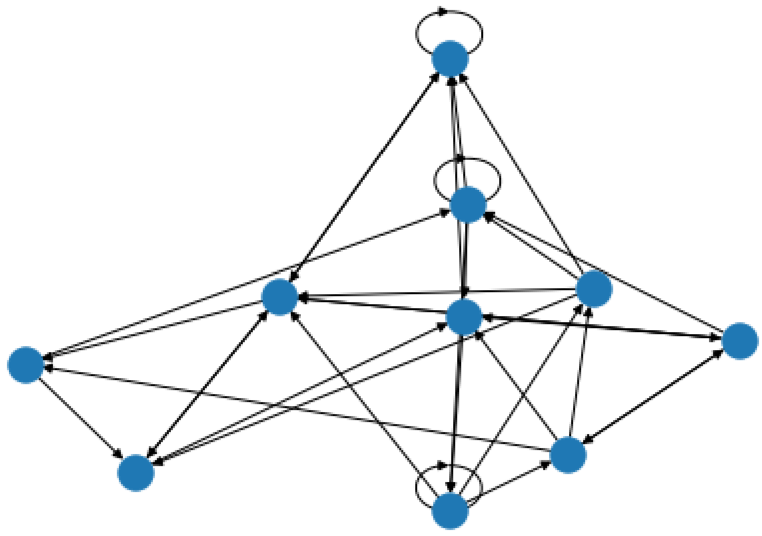
Build a graph

In [ ]:

```
import networkx as nx
import numpy as np
import torch
```

In [ ]:

```
G1 = nx.read_edgelist('',create_using=nx.DiGraph(),nodetype=None,data=[('weig
nx.draw(G1)
```



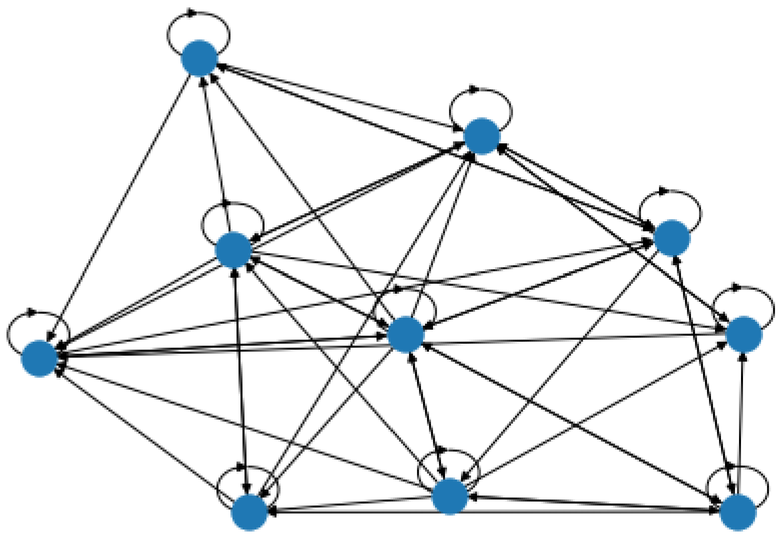
In [ ]:

```python
A_torch = torch.randint(0, 2, (10, 10))
A_np = np.random.randint(2, size=(15, 15))

def build_graph(adj_matrix):
    graph = nx.DiGraph()  # note nx.Graph() v.s. nx.DiGraph()
    for i in range(len(adj_matrix)):
        for j in range(len(adj_matrix[i])):
            if adj_matrix[i][j] == 1:
                graph.add_edge(i, j)
            if i == j:
                graph.add_edge(i, j)
    return graph
```

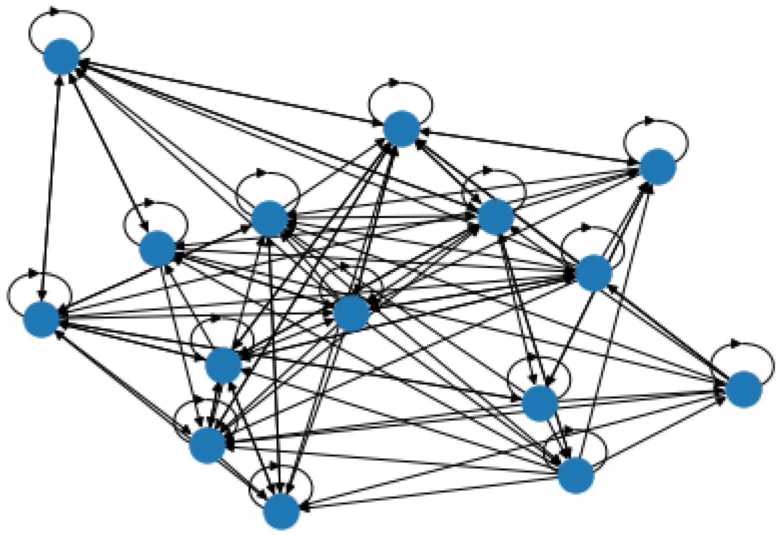
In [ ]:

```python
G2 = build_graph(A_torch)
nx.draw(G2)
```



In [ ]:

```python
G3 = build_graph(A_np)
nx.draw(G3)
```



Implement DeepWalk

Packages: Import necessary packages

In [ ]:

```
import networkx as nx
from joblib import Parallel, delayed
import random
import itertools
import numpy as np
from gensim.models import Word2Vec
```

Utils: Processing data

In [ ]:

```
def partition_num(num, workers):
    if num % workers == 0:
        return [num//workers]*workers
    else:
        return [num//workers]*workers + [num % workers]
```

Model: DeepWalk

Algorithm 1 DEEPWALK(G, w, d, γ, t)

Input: graph $G(V, E)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$
1: Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$
2: Build a binary Tree T from V
3: **for** $i = 0$ to γ **do**
4: $\quad \mathcal{O} = \text{Shuffle}(V)$
5: $\quad$ **for each** $v_i \in \mathcal{O}$ **do**
6: $\quad\quad \mathcal{W}_{v_i} = RandomWalk(G, v_i, t)$
7: $\quad\quad$ SkipGram(Φ, $\mathcal{W}_{v_i}$, w)
8: $\quad$ **end for**
9: **end for**

In [ ]:

```
class RandomWalker:
  def __init__(self, G, num_walks, walk_length):
      """
      :param G: Graph
      :param num_walks: a number of walks per vertex
      :param walk_length: Length of a walk. Each walk is considered as a sent
      """
      self.G = G
      self.num_walks = num_walks
      self.walk_length = walk_length


  def deepwalk_walk(self, start_node):
      """
      :param start_node: Starting node of a walk
      """
      walk = [start_node]
      while len(walk) < self.walk_length:
          cur = walk[-1]
          # Check if having any neighbors at the current node
          cur_nbrs = list(self.G.neighbors(cur))
          if len(cur_nbrs) > 0:
              # Random walk with the probability of 1/d(v^t). d(v^t) is the n
              walk.append(random.choice(cur_nbrs))
          else:
              break
      return walk


  def simulate_walks(self, workers=1, verbose=0):
      """
      :param workers: a number of workers running in parallel processing
      :param verbose: progress bar
      """
      G = self.G
      nodes = list(G.nodes())
      results = Parallel(n_jobs=workers, verbose=verbose)(
          delayed(self._simulate_walks)(nodes) for num in
          partition_num(self.num_walks, workers))
      walks = list(itertools.chain(*results))
      return walks

  # INFORMATION EXTRACTOR
```

```python
    def _simulate_walks(self, nodes):
        walks = []
        # Iterate all walks per vertex
        for _ in range(self.num_walks):
            random.shuffle(nodes)
            # Iterate all nodes in a walk
            for v in nodes:
              walks.append(self.deepwalk_walk(start_node=v))
        return walks
```

In [ ]:

```python
class DeepWalk:
    def __init__(self, graph, walk_length, num_walks, workers=1):

        self.graph = graph
        self.w2v_model = None
        self._embeddings = {}

        self.walker = RandomWalker(graph, num_walks=num_walks, walk_length=wa
        self.sentences = self.walker.simulate_walks(workers=workers, verbose=

    def train(self, embed_size=128, window_size=5, workers=1, iter=5, **kwarg

        kwargs["sentences"] = self.sentences
        kwargs["min_count"] = kwargs.get("min_count", 0)
        kwargs["size"] = embed_size
        kwargs["sg"] = 1  # skip gram
        kwargs["hs"] = 1  # deepwalk use Hierarchical Softmax
        kwargs["workers"] = workers
        kwargs["window"] = window_size
        kwargs["iter"] = iter

        print("Learning embedding vectors...")
        model = Word2Vec(**kwargs) # Pay attention here
        print("Learning embedding vectors done!")

        self.w2v_model = model
        return model

    def get_embeddings(self,):
        if self.w2v_model is None:
            print("model not train")
            return {}

        self._embeddings = {}
        for word in self.graph.nodes():
            self._embeddings[word] = self.w2v_model.wv[word]

        return self._embeddings
```

Run graph embedding

In [ ]:

```python
G = nx.read_edgelist('',create_using=nx.DiGraph(),nodetype=None,data=[('weigh
model = DeepWalk(G,walk_length=10,num_walks=80,workers=1)#init model
model.train(window_size=5,iter=3)# train model
embeddings = model.get_embeddings()# get embedding vectors
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent worker
s.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    4.8s finished
Learning embedding vectors...
Learning embedding vectors done!
```

In [ ]:

```python
count = 0
for i, (k, v) in enumerate(embeddings.items()):
  print("Index {} has key {} and value {}".format(str(i), k, v))
  count += 1
  if count == 2:
    break
```

```
Index 0 has key 1397 and value [ 0.34124643  0.21651596  0.18768169  0.0116240
6  0.31037042 -0.21311687
 -0.21127611 -0.11157233 -0.29168016 -0.16498274 -0.56216705  0.7877994
  0.16568734 -0.16140895 -0.04564727  0.5001673   0.46898863  0.03051835
  0.00187808 -0.71714365  0.29957256 -0.30319893 -0.31948227 -0.1334579
 -0.11667724  0.07388762  0.33723173  0.29406253  0.10739747  0.09178717
 -0.6103274   0.09980148  0.08364724  0.03909857 -0.37748447 -0.34338912
  0.24248604  0.11645296  0.41313785 -0.03004063  0.8495428   0.42406577
 -0.22229639  0.14524612 -0.2500054  -0.0770914   0.04825141 -0.40773597
 -0.31535515 -0.23247224 -0.5798591  -0.5841959  -0.2367024  -0.09526283
  0.15386958 -0.06518534 -0.04644093 -0.258978   -0.5367063  -0.03583927
  0.69152796 -0.1910452   0.04655356 -0.04441753 -0.60192573  0.2539498
  0.26476353  0.05972561  0.04416491 -0.34739408 -0.06243756  0.21690315
  0.3327238  -0.05461631  0.05624618  0.28194264  0.19865002  0.14220454
  0.06796283 -0.7308261   0.27937174  0.18426932 -0.60932744 -0.40930104
 -0.41955045  0.07619184 -0.23528285 -0.08905817 -0.28653207 -0.0128131
  0.08336076  0.058368    0.04754102 -0.0565265   0.02882961 -0.34405535
  0.45909476  0.450402   -0.1676697   0.35118464  0.15465385 -0.21801348
  0.31429133 -0.40304697 -0.32297945  0.23116153  0.605175    0.3690839
  0.05721382 -0.20259163 -0.72533935  0.05584092  0.63240725 -0.03199526
 -0.624229   -0.41470948 -0.01885439 -0.09232835  0.05836875  0.4564771
  0.4306216  -0.50073427 -0.11824699  0.1369     -0.16162194  0.02561174
 -0.23480387  0.30287743]
Index 1 has key 1470 and value [-9.85323451e-04  2.44411126e-01 -1.11971118e-0
1  1.20491803e-01
 -4.53836098e-02 -2.37610996e-01  9.34034213e-02 -3.91296387e-01
 -1.88627854e-01 -1.29437178e-01 -3.16280633e-01  3.21089029e-01
  4.66137260e-01  3.61894518e-02  1.35226890e-01  2.84080923e-01
  7.58473337e-01  1.56206280e-01  9.57813580e-03 -6.69257462e-01
  2.21789196e-01  1.09778130e-02 -2.16602907e-01 -7.85740092e-02
  1.16624899e-01 -4.13993984e-01  1.27028331e-01 -2.07150623e-01
 -4.08071429e-02  6.69466794e-01 -2.50516593e-01 -1.03555061e-02
  1.30140513e-01 -1.69554248e-01  1.72479197e-01 -8.22005495e-02
  4.75637227e-01  1.95431188e-01  1.99893594e-01  4.38632876e-01
  3.73338789e-01  6.32987842e-02  3.23832244e-01  2.27926120e-01
 -5.65705240e-01 -1.89654574e-01 -1.48081392e-01 -8.05980042e-02
 -4.94810700e-01 -4.86125052e-01 -5.98641515e-01 -8.10829043e-01
 -1.41557649e-01 -4.20798004e-01  2.40260705e-01 -2.11428002e-01
 -3.61233175e-01 -3.60062003e-01 -5.70150971e-01  2.08269730e-01
  6.37328684e-01  1.05934627e-01 -3.59246612e-01  7.80420601e-02
 -3.61679286e-01  1.32018313e-01  5.27841747e-01  8.48537758e-02
  3.04608047e-01  5.02530485e-02 -1.46592200e-01  3.24253172e-01
  2.66865432e-01 -7.99008533e-02  8.87226611e-02 -4.12208550e-02
  6.70765340e-02  6.06886446e-01 -2.53769644e-02 -7.92414963e-01
  1.75212950e-01  4.88836259e-01 -7.49678254e-01 -5.02590239e-01
 -7.17394173e-01  2.54527688e-01 -4.59037811e-01 -3.68008673e-01
 -1.21800052e-02 -1.72954984e-02 -1.15540083e-02 -1.05653279e-01
  2.47311946e-02  1.50966838e-01 -1.96573250e-02 -5.39956450e-01
  1.01292014e+00  4.69347060e-01 -1.70098901e-01  3.13794911e-01
  1.66759714e-01 -1.26604021e-01  4.64111388e-01 -7.92570472e-01
 -1.59762695e-01  4.52025443e-01  6.53884649e-01  8.37768242e-02
 -1.22993879e-01  4.08139676e-02 -5.62696755e-01  1.27226844e-01
  2.68604755e-01 -3.38216335e-01 -4.96307015e-01 -5.26855171e-01
 -2.71428257e-01 -3.47952634e-01 -1.94664687e-01  4.14689809e-01
  2.13804364e-01 -7.32546210e-01  2.99578696e-01  2.48458609e-01
 -1.15290515e-01 -8.36742595e-02  5.82921132e-02  5.65196097e-01]
```

Questions

Did you see that we use the function "Word2vec" as the primary function to implement the DeepWalk algorithm?

The reason is that DeepWalk is based on the idea of Word2vec. As a result, all we need is packed in the implementation of Word2vec. Within a short amount of time, we couldn't go through all the code.

This is your homework. The details will be shown in the file "Lab3 - Homeworks".

Please take a look at this file for more details.