

Supplementary Materials to A Comprehensive Survey on Graph Neural Networks

Zonghan Wu, Shirui Pan, *Member, IEEE*, Fengwen Chen, Guodong Long,
Chengqi Zhang, *Senior Member, IEEE*, Philip S. Yu, *Fellow, IEEE*

A. Data Set

Citation Networks consist of papers, authors, and their relationships such as citations, authorship, and co-authorship. Although citation networks are directed graphs, they are often treated as undirected graphs in evaluating model performance with respect to node classification, link prediction, and node clustering tasks. There are three popular data sets for paper-citation networks, Cora, Citeseer and Pubmed. The Cora data set contains 2708 machine learning publications grouped into seven classes. The Citeseer data set contains 3327 scientific papers grouped into six classes. Each paper in Cora and Citeseer is represented by a one-hot vector indicating the presence or absence of a word from a dictionary. The Pubmed data set contains 19717 diabetes-related publications. Each paper in Pubmed is represented by a term frequency-inverse document frequency (TF-IDF) vector. Furthermore, DBLP is a large citation data set with millions of papers and authors which are collected from computer science bibliographies. The raw data set of DBLP can be found on <https://dblp.uni-trier.de>. A processed version of the DBLP paper-citation network is updated continuously by <https://aminer.org/citation>.

Biochemical Graphs Chemical molecules and compounds can be represented by chemical graphs with atoms as nodes and chemical bonds as edges. This category of graphs is often used to evaluate graph classification performance. The NCI-1 and NCI-9 data set contain 4110 and 4127 chemical compounds respectively, labeled as to whether they are active to hinder the growth of human cancer cell lines. The MUTAG data set contains 188 nitro compounds, labeled as to whether they are aromatic or heteroaromatic. The D&D and PROTEIN data set represent proteins as graphs, labeled as to whether they are enzymes or non-enzymes. The PTC data set consists of 344 chemical compounds, labeled as to whether they are carcinogenic for male and female rats. The QM9 data set records 13 physical properties of 133885 molecules with up to 9 heavy atoms. The Alchemy data set records 12 quantum mechanical properties of 119487 molecules comprising up to 14 heavy atoms. Another important data set is the Protein-Protein Interaction network (PPI). It contains 24 biological graphs with nodes represented by proteins and edges represented by the interactions between proteins. In PPI, each graph is associated with one human tissue. Each node is labeled with its biological states.

Social Networks are formed by user interactions from online services such as BlogCatalog and Reddit. The BlogCatalog data set is a social network which consists of bloggers and their social relationships. The classes of bloggers represent their personal interests. The Reddit data set is an undirected graph formed by posts collected from the Reddit discussion forum. Two posts are linked if they contain comments by the same user. Each post has a label indicating the community to which it belongs.

Others There are several other data sets worth mentioning. The MNIST data set contains 70000 images of size 28×28 labeled with ten digits. An MNIST image is converted to a graph by constructing an 8-nearest-neighbors graph based on its pixel locations. The METR-LA is a spatial-temporal graph data set. It contains four months of traffic data collected by 207 sensors on the highways of Los Angeles County. The adjacency matrix of the graph is computed by the sensor network distance with

TABLE I: Reported experimental results for node classification on five frequently used data sets. Cora, Citeseer, and Pubmed are evaluated by classification accuracy. PPI and Reddit are evaluated by micro-averaged F1 score.

Method	Cora	Citeseer	Pubmed	PPI	Reddit
SSE (2018)	-	-	-	83.60	-
GCN (2016)	81.50	70.30	79.00	-	-
Caylennets (2017)	81.90	-	-	-	-
DualGCN (2018)	83.50	72.60	80.00	-	-
GraphSage (2017)	-	-	-	61.20	95.40
GAT (2017)	83.00	72.50	79.00	97.30	-
MoNet (2017)	81.69	-	78.81	-	-
LGCN (2018)	83.30	73.00	79.50	77.20	-
GAAN (2018)	-	-	-	98.71	96.83
FastGCN (2018)	-	-	-	-	93.70
StoGCN (2018)	82.00	70.90	78.70	97.80	96.30
Huang et al. (2018)	-	-	-	-	96.27
GeniePath (2019)	-	-	78.50	97.90	-
DGI (2018)	82.30	71.80	76.80	63.80	94.00
Cluster-GCN (2019)	-	-	-	99.36	96.60

TABLE II: A Summary of Open-source Implementations

Model	Framework	Github Link
GGNN (2015)	torch	https://github.com/yujiali/ggnn
SSE (2018)	c	https://github.com/Hanjun-Dai/steady_state_embedding
ChebNet (2016)	tensorflow	https://github.com/mdeff/cnn_graph
GCN (2017)	tensorflow	https://github.com/tkipf/gcn
CayleyNet (2017)	tensorflow	https://github.com/amoliu/CayleyNet
DualGCN (2018)	theano	https://github.com/ZhuangCY/DGCN
GraphSage (2017)	tensorflow	https://github.com/williamleif/GraphSAGE
GAT (2017)	tensorflow	https://github.com/PetarV-/GAT
LGCN (2018)	tensorflow	https://github.com/divelab/lgcn/
PGC-DGCNN (2018)	pytorch	https://github.com/dinhinfotech/PGC-DGCNN
FastGCN (2018)	tensorflow	https://github.com/matenure/FastGCN
StoGCN (2018)	tensorflow	https://github.com/thu-ml/stochastic_gcn
DGCNN (2018)	torch	https://github.com/muhanzhang/DGCNN
DiffPool (2018)	pytorch	https://github.com/RexYing/diffpool
DGI (2019)	pytorch	https://github.com/PetarV-/DGI
GIN (2019)	pytorch	https://github.com/weihua916/powerful-gnns
Cluster-GCN (2019)	pytorch	https://github.com/benedekrozemberczki/ClusterGCN
DNGR (2016)	matlab	https://github.com/ShelsonCao/DNGR
SDNE (2016)	tensorflow	https://github.com/suanrong/SDNE
GAE (2016)	tensorflow	https://github.com/limaosen0/Variational-Graph-Auto-Encoders
ARVGA (2018)	tensorflow	https://github.com/Ruiqi-Hu/ARGA
DRNE (2016)	tensorflow	https://github.com/tadpole/DRNE
GraphRNN (2018)	tensorflow	https://github.com/snap-stanford/GraphRNN
MolGAN (2018)	tensorflow	https://github.com/nicola-decao/MolGAN
NetGAN (2018)	tensorflow	https://github.com/danielzuegner/netgan
GCRN (2016)	tensorflow	https://github.com/youngjoo-epfl/gconvRNN
DCRNN (2018)	tensorflow	https://github.com/liyaguang/DCRNN
Structural RNN (2016)	theano	https://github.com/asheshjain399/RNNexp
CGCN (2017)	tensorflow	https://github.com/VeritasYin/STGCN_IJCAI-18
ST-GCN (2018)	pytorch	https://github.com/yysijie/st-gcn
GraphWaveNet (2019)	pytorch	https://github.com/nnzhan/Graph-WaveNet
ASTGCN (2019)	mxnet	https://github.com/Davidham3/ASTGCN

a Gaussian threshold. The NELL data set is a knowledge graph obtained from the Never-Ending Language Learning project. It consists of facts represented by a triplet which involves two entities and their relation.

B. Reported Experimental Results for Node Classification

A summarization of experimental results of methods which follow a standard train/valid/test split is given in Table I.

C. Open-source Implementations

Here we summarize the open-source implementations of graph neural networks reviewed in the survey. We provide the hyperlinks of the source codes of the GNN models in table II.