1. What is citation network?

Wikipedia

Citation Network is a social network which contains paper sources and linked by co-citation relationships. Egghe & Rousseau once (1990, p. 228) explain "when a document d_i cites a document d_j, we can show this by an arrow going from the node representing d_i to the document representing d_j. In this way the documents from a collection D form a directed graph, which is called a 'citation graph' or 'citation network' "

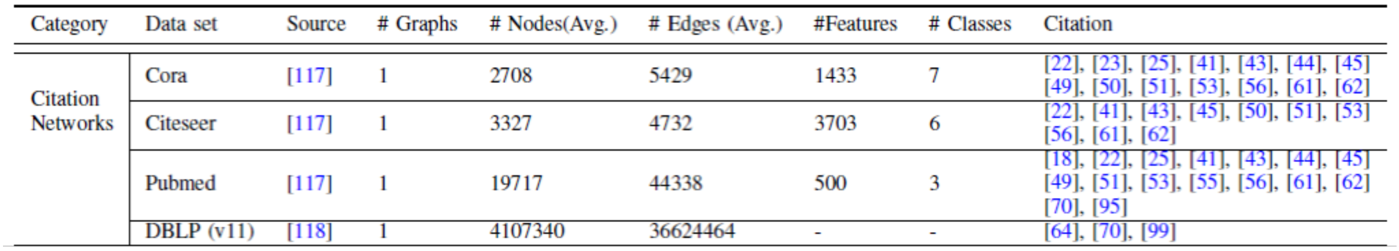
2. Data sets

2019 “A Comprehensive Survey on Graph Neural Networks”

APPENDIX A.

Data Set

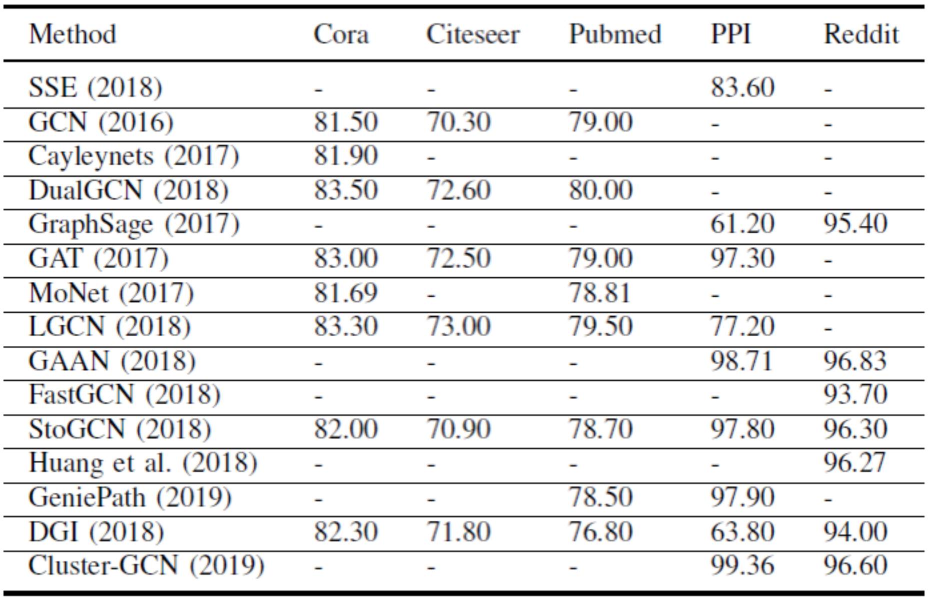
TABLE VI: Summary of selected benchmark data sets.



Category	Data set	Source	# Graphs	# Nodes(Avg.)	# Edges (Avg.)	#Features	# Classes	Citation
Citation Networks	Cora	[117]	1	2708	5429	1433	7	[22], [23], [25], [41], [43], [44], [45] [49], [50], [51], [53], [56], [61], [62]
	Citeseer	[117]	1	3327	4732	3703	6	[22], [41], [43], [45], [50], [51], [53] [56], [61], [62]
	Pubmed	[117]	1	19717	44338	500	3	[18], [22], [25], [41], [43], [44], [45] [49], [51], [53], [55], [56], [61], [62] [70], [95]
	DBLP (v11)	[118]	1	4107340	36624464	-	-	[64], [70], [99]

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 papercitation 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.

TABLE VII: 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	-	-
Cayleynets (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

3. How to train a GNN model?

2017 “Semi-Supervised Classification with Graph Convolutional Networks”

5 Experiments

We test our model in a number of experiments: semi-supervised document classification in citation networks, semi-supervised entity classification in a bipartite graph extracted from a knowledge graph, an evaluation of various graph propagation models and a run-time analysis on random graphs.

5.1 Datasets

We closely follow the experimental setup in Yang et al. (2016). Dataset statistics are summarized in Table 1. In the citation network datasets—Citeseer, Cora and Pubmed (Sen et al., 2008)—nodes are documents and edges are citation links. Label rate denotes the number of labeled nodes that are used for training divided by the total number of nodes in each dataset. NELL (Carlson et al., 2010; Yang et al., 2016) is a bipartite graph dataset extracted from a knowledge graph with 55,864 relation nodes and 9,891 entity nodes.

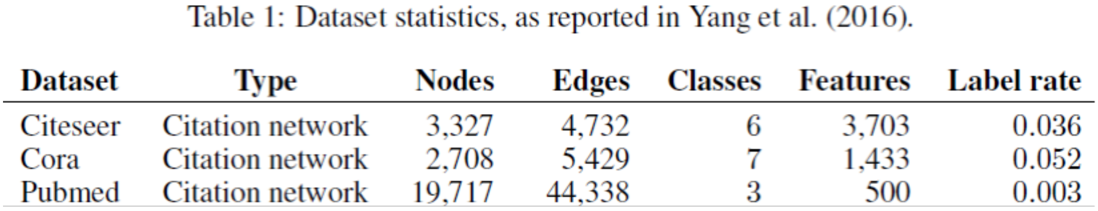
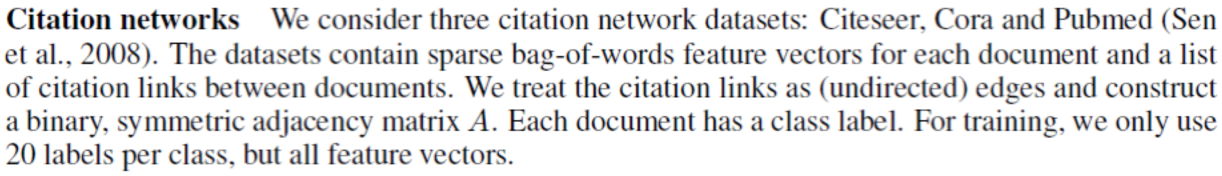


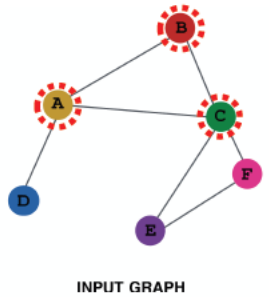
Table 1: Dataset statistics, as reported in Yang et al. (2016).

Dataset	**Type**	**Nodes**	**Edges**	**Classes**	**Features**	**Label rate**
Citeseer	Citation network	3,327	4,732	6	3,703	0.036
Cora	Citation network	2,708	5,429	7	1,433	0.052
Pubmed	Citation network	19,717	44,338	3	500	0.003



Citation networks We consider three citation network datasets: Citeseer, Cora and Pubmed (Sen et al., 2008). The datasets contain sparse bag-of-words feature vectors for each document and a list of citation links between documents. We treat the citation links as (undirected) edges and construct a binary, symmetric adjacency matrix A. Each document has a class label. For training, we only use 20 labels per class, but all feature vectors.

Overview of Model Design



3) Train on a set of nodes, i.e., a batch of compute graphs