Table 3: F1-score on Cora dataset, where the percentage of labeled nodes varies from 10% to 50%.

% Labeled Nodes	10%	20%	30%	40%	50%
DeepWalk	76.4	78.0	79.5	80.5	81.0
MMDW	74.9	80.8	82.8	83.7	84.7
Text Features	58.3	67.4	71.1	73.3	74.0
PLSA	57.0	63.1	65.1	66.6	67.6
Naive Combination	76.5	80.4	82.3	83.3	84.1
NetPLSA	80.2	83.0	84.0	84.9	85.4
TADW	82.4	85.0	85.6	86.0	86.7
STNE	**84.2**	**86.5**	**87.0**	**86.9**	**88.2**

Table 4: F1-score on Citeseer dataset, where the percentage of labeled nodes varies from 10% to 50%.

% Labeled Nodes	10%	20%	30%	40%	50%
DeepWalk	52.4	54.7	56.0	56.5	57.3
MMDW	55.6	60.1	63.2	65.1	66.9
Text Features	58.3	66.4	69.2	71.2	72.2
PLSA	54.1	58.3	60.9	62.1	62.6
Naive Combination	61.0	66.7	69.1	70.8	72.0
NetPLSA	58.7	61.6	63.3	64.0	64.7
TADW	**70.6**	**71.9**	**73.3**	73.7	74.2
STNE	69.6	71.2	72.2	**74.3**	**74.8**

Table 5: F1-score on Wiki dataset, where the percentage of labeled nodes varies from 10% to 50%.

% Labeled Nodes	10%	20%	30%	40%	50%
DeepWalk	59.3	64.3	66.2	68.1	68.8
MMDW	57.8	62.3	65.8	67.3	67.3
Text Features	65.1	72.9	75.6	77.1	77.4
PLSA	69.0	72.5	74.7	75.5	76.0
Naive Combination	66.3	73.0	75.2	77.1	78.6
NetPLSA	67.2	70.6	71.7	71.9	72.3
TADW	72.6	77.3	79.2	79.9	80.3
STNE	**73.9**	**78.0**	**80.6**	**81.5**	**82.7**

makes the content information relatively sparse. Second, documents in Citeseer datasets contain more words than those in Cora dataset, but there are fewer edges among nodes. So the content information contributes relatively more to performances on Citeseer dataset. Third, although the Wiki dataset contains far more edges than the other two, the long documents seem to contribute more to the classification. However, by simultaneously modeling content and structure information, Naive, NetPLSA, TADW and STNE can all outperform the content only or structure only baselines.

- On Cora and Wiki datasets, STNE outperforms all compared baselines. On Citeseer dataset, STNE achieves comparable results to TADW, but still outperforms other baselines. These observations can demonstrate the effectiveness of STNE. As Table 1 shows, Citeseer dataset has more nodes but fewer edges than the other two, thus the edge density is lower. Because our STNE utilizes higher order proximity information than TADW, it suffers more from the insufficient links.
- On Cora and Wiki datasets, STNE can outperform compared baselines even when fewer nodes are labeled. The F1-scores of most compared baselines would drop when the classifier is trained with fewer labeled nodes. Reason is that they cannot effectively combine the content and structure information, and inconsistencies exist between the representations of training and testing samples. Because STNE can jointly learn from content and structure information, and take advantage of long node sequences to smooth nodes' representations, the training and testing samples would be more consistent.

5.5 Parameter Analysis

There are two important parameters in STNE, the length of random walks and the number of walks started at each node. In this subsection, we evaluate how different values of walk length and walk number can affect the results.

5.5.1 Length of Random Walks. We first show how the length of random walks affects the performance in Figure 3, where it varies from 3 to 15. We can see that initially the performance raises a little when the length of random walks increases. This is intuitive as longer sequences can encode higher order proximity information. However, when the length of walks continuously increases after it reaches 10, the performance starts to drop slowly. The reason is that too long sequences may introduce noises and deteriorate performances. Compared to Figure 5 in node2vec [9], STNE is less sensitive to the length of random walks. Thus jointly modeling both content and structure information can reduce the dependence on high order proximity information. In addition, STNE only requires length 10 walks to achieve best performances, while DeepWalk and node2vec require 40.

5.5.2 Number of Random Walks per Node. Then we show how the number of random walks started at each node affects the performance in Figure 4, where it varies from 1 to 15. The number of random walks decides how many times we traverse the inner sequential structure information contained in graphs. As can be observed in Figure 4, a small number of random walks is incapable of fully exploiting the structure information, and performance would be improved when it increases. However, when the number of walks continuously increases after it reaches 10, the performance starts to drop slowly. The reason is that too many random walks may introduce noises and redundancy and deteriorate performance. In addition, compared to Figure 5 in DeepWalk [19] and Figure 5 in node2vec [9], STNE is less sensitive to the number of random walks, as the content information can help reduce the dependence on structure information.

5.6 Node Representation Analysis

Besides the encoder output $\mathbf{h}(v_i)$, the decoder output $\mathbf{d}(v_i)$ can also be similarly calculated as in Equation (17). Both $\mathbf{h}$s and $\mathbf{d}$s can be taken as node representations in subsequent tasks, so as the concatenation of them. It is interesting to investigate the performances of different node representations in classification. The F1-scores are plotted out in Figure 5, where the percent of labeled nodes r varies from 10% to 50%. The overall trend is similar on all three datasets, where encoder output $\mathbf{h}$ achieves the best F1-scores,