A GRAPH CONSTRUCTION STUDY FOR GRAPH-BASED SSL

CASE STUDY OF DISTANCE METRICS ON UNSTRUCTURED DATA

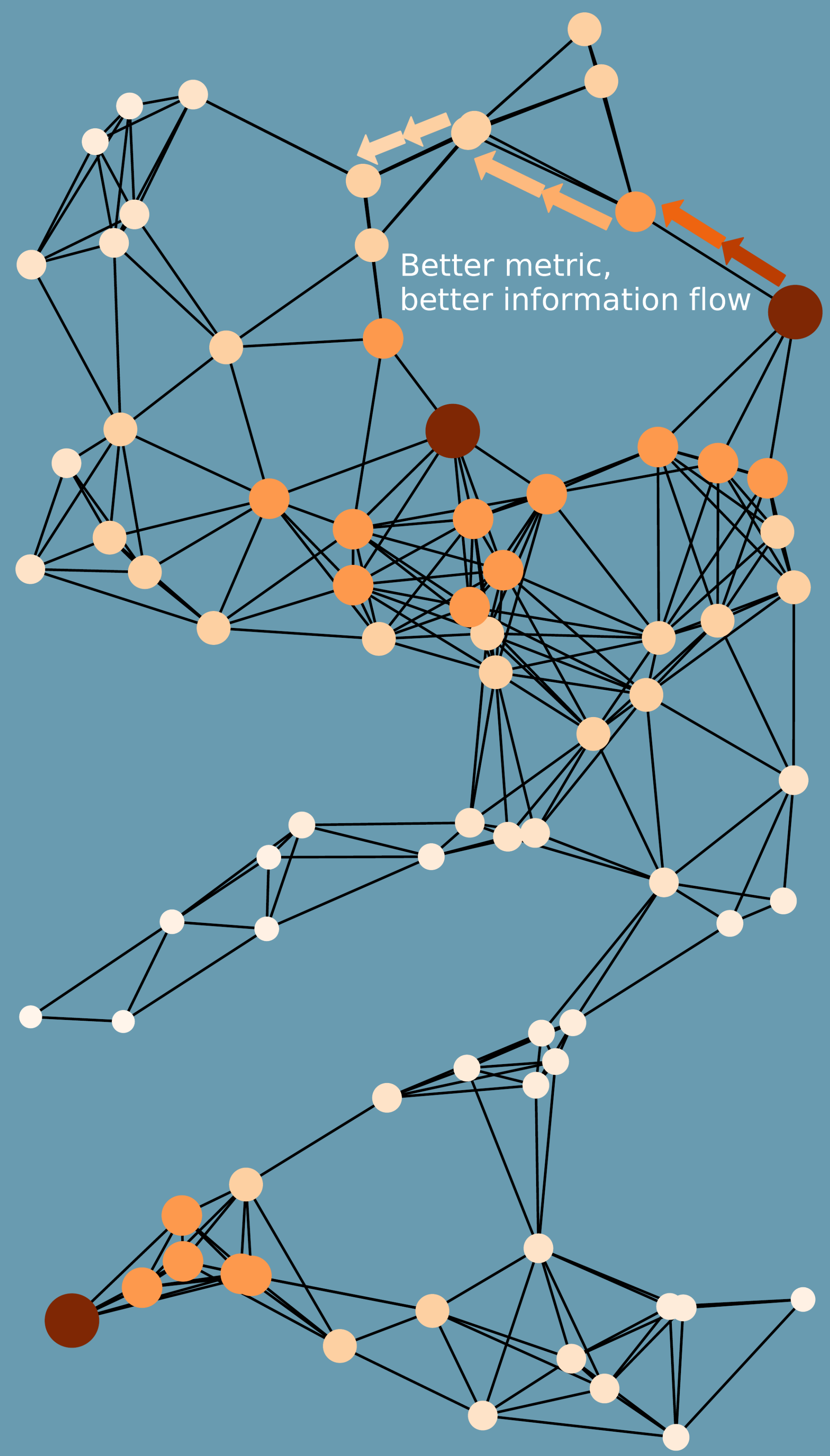
Sumedh Yadav, Gautam Kumar, Shivam Kumar*[1]

[1]*Gstech Tech. Pvt. Ltd., Bengaluru, India*

#ODSC, LONDON, 2019

SIGNIFICANCE

- Similarity metrics are decisively better than L2.
- Implementation of two similarity metrics in the fast library for approximate nearest neighbors (FLANN).
- Construction of knowledge graph with text data in mind.



INTERESTED?

Project webpage

More projects in ML, Big Data

CONTACT

Sumedh is an application data scientist, and has been working in Indian E-commerce market for four years. He recently developed algorithms for distributed machine learning and *fairpe* product.

sumedhyadav.iitkgp@gmail.com Bengaluru

1. LEAD-IN

- Similarity metric cosine similarity (CS) and particularly state-of-the-art improved sqrt-cosine (ISC) similarity are shown to be effective (Sohangir et al. 2017) for text data.
- Graph-based SSL algorithms are traditionally popular among graph structured datasets (Subramanya et al. 2014), however, graphs can be used to represent data in an organic way (Wu et al. 2018).
- We present a study on how the distance/similarity metrics impact the graph construction and the subsequent classification task...

***Keywords** - graph construction, similarity metrics, unstructured text data, graph-based SSL*

3. DATASETS

Two text datasets are considered,

- A E-comm. dataset, Dataset I, depicted in Figure 2.

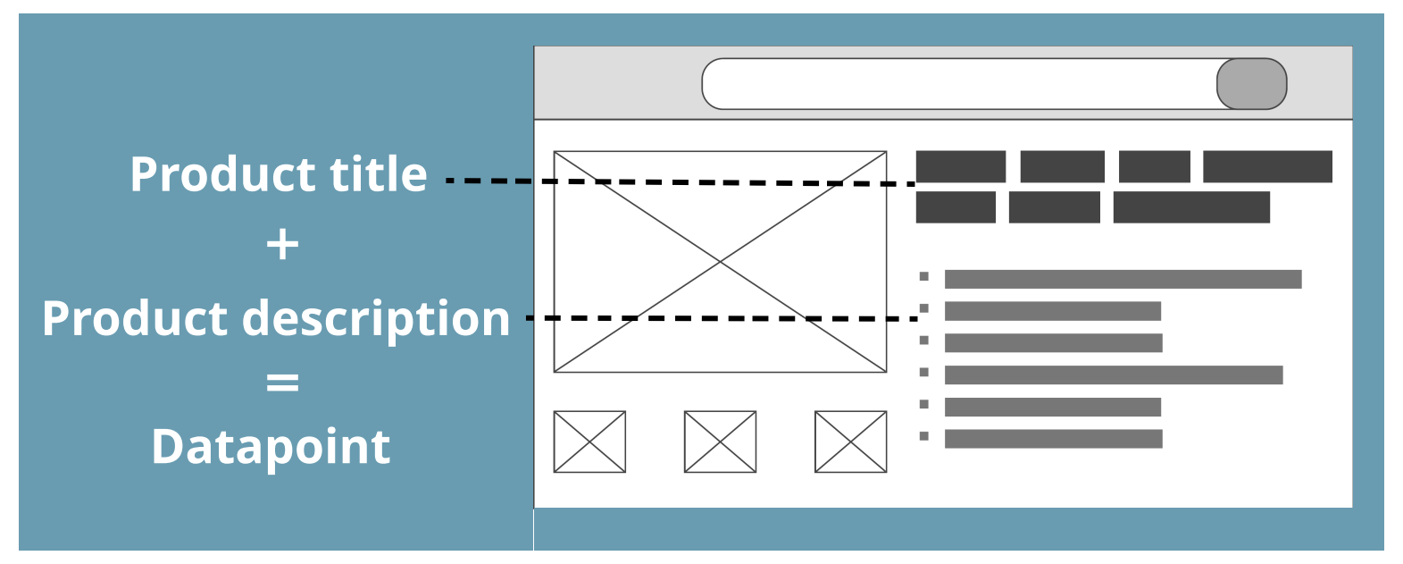
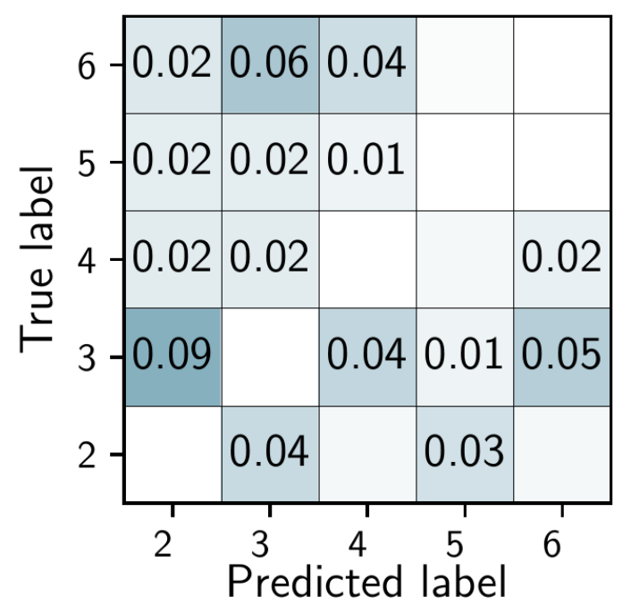


Figure 2 A typical product page on a E-commerce platform has plenty of text data.

- 20 newsgroups data, Dataset II, for which Table 1 was used to choose the four most confused classes.

Table 1 Partial truth table of Dataset II.



True label \ Predicted label	2	3	4	5	6
6	0.02	0.06	0.04		
5	0.02	0.02	0.01		
4	0.02	0.02			0.02
3	0.09		0.04	0.01	0.05
2		0.04		0.03	

Feature engineering is key for text data, and steps shown in Figure 3 were involved.

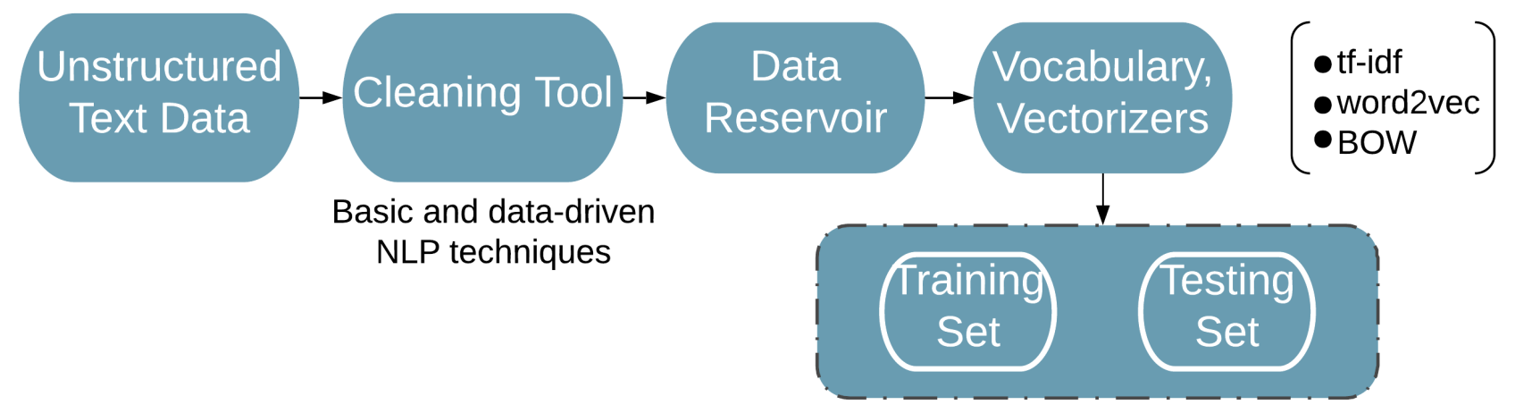


Figure 3 Feature engineering pipeline before graph construction.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Andreas Loukas for his suggestion related to graph-based SSL. Furthermore, we express our thanks to Mr. Shreyas Puttige for his input related to the poster design.

WHAT'S NEW?

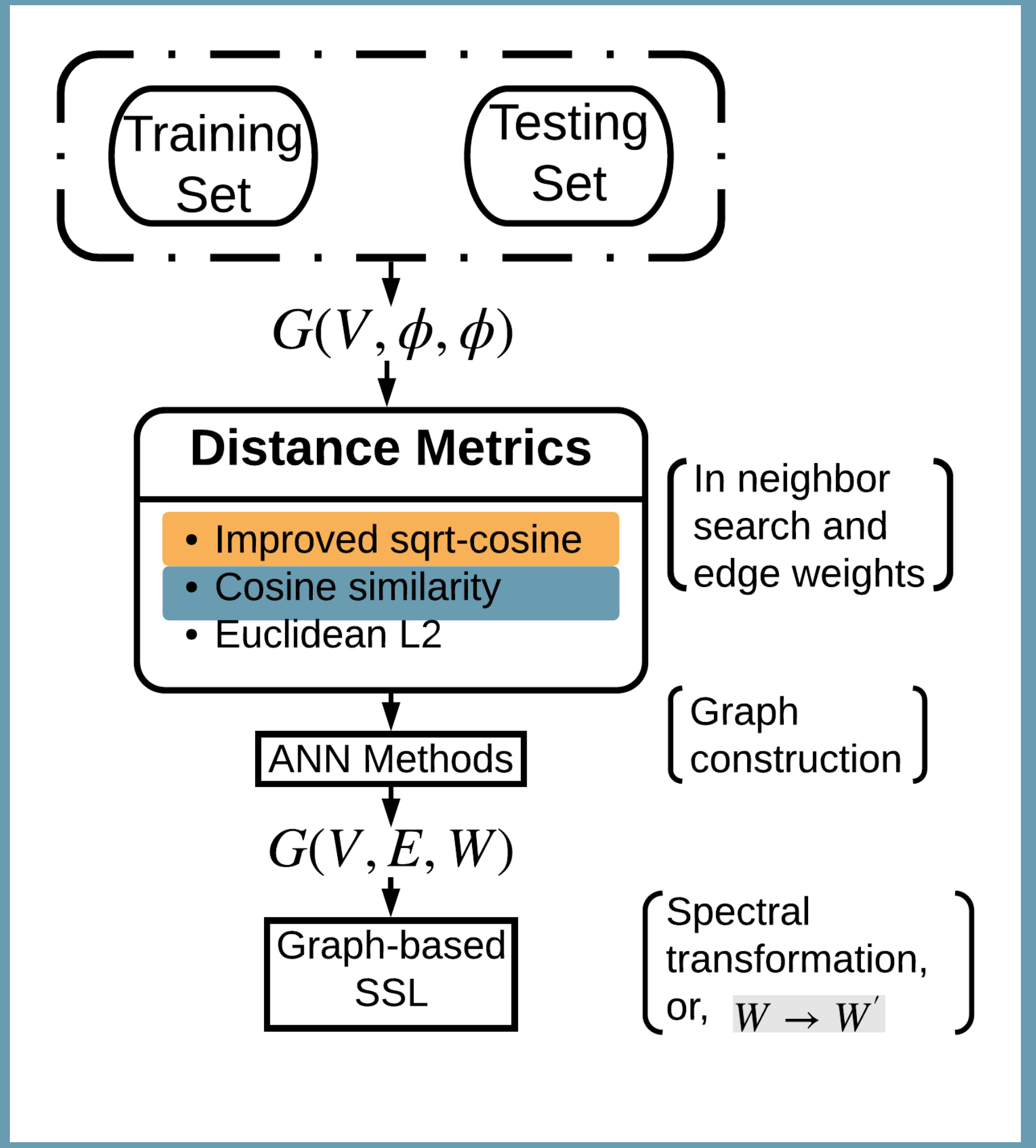


Figure 1 Graph construction and transductive classification procedure with new implementations highlighted in color

WHICH IS THE BEST?

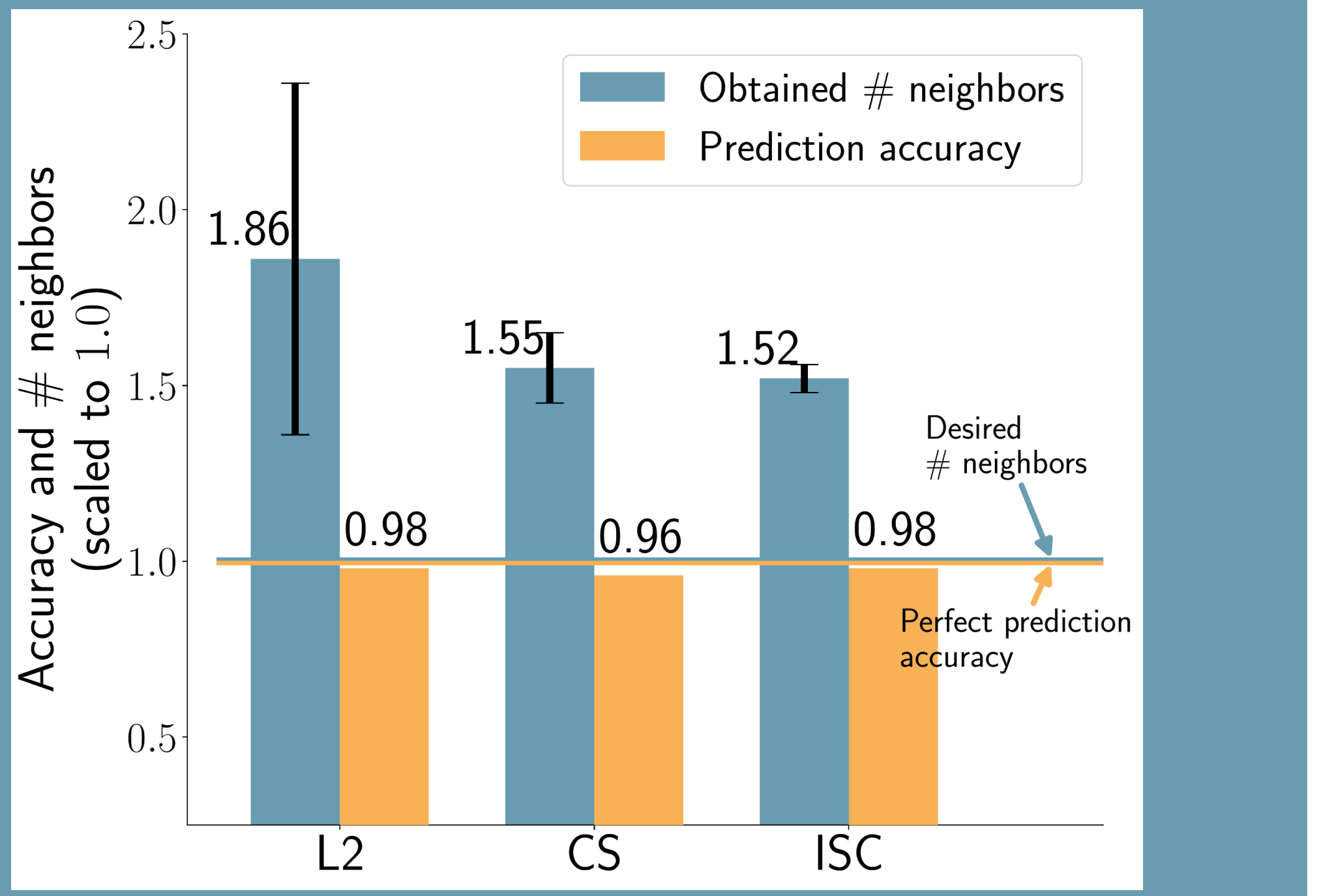


Figure 6 On Dataset I, ISC performs best in graph construction.

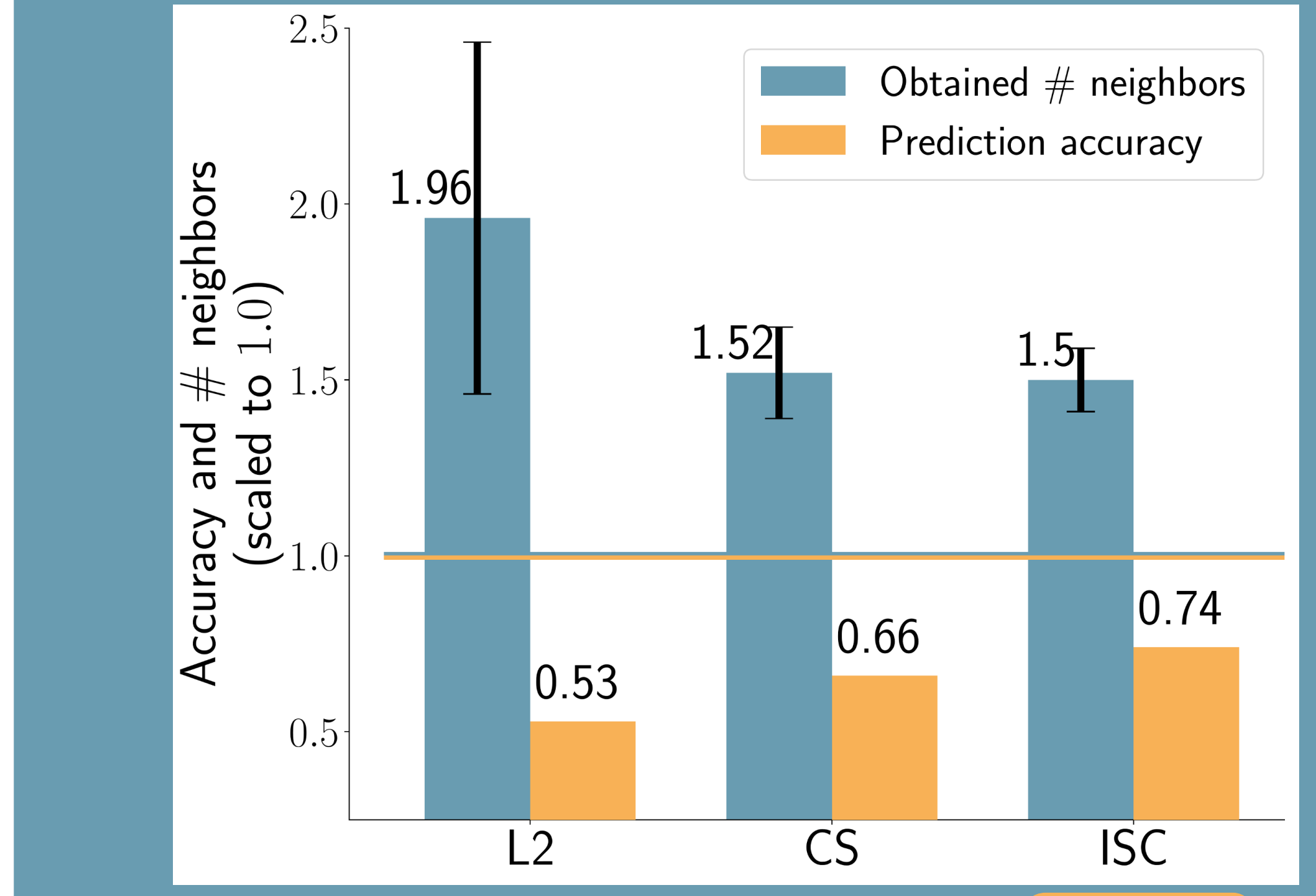


Figure 7 On Dataset II, a clear performance trend of ISC > CS > L2 is observed.

2. THE MATHEMATICS

- Graph construction, as shown in Figure 1, uses the similarity metrics,

$$\mathrm{ISC}(x,y) = \frac{\sum_{i=1}^{n} \sqrt{x_i y_i}}{\sqrt{\sum_{i=1}^{n} x_i}\sqrt{\sum_{i=1}^{n} y_i}} \quad \& \quad \mathrm{CS}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

  where x and y are nodes and n is the number of features.
- Algorithm 1 outlines the graph construction procedure using FLANN library.

```
Algorithm 1: Graph construction
  Input: $G(V, \phi, \phi)$, $nn$ - number of neighbors,
         $metric$ - ISC/CS/L2
  Output: $G(V, E, W)$

$flann() \leftarrow V$                          ▷ space search is formed
$raw\_E \leftarrow flann.ann\_search(nn, metric)$
for $i \in V$ do
    for $j \in raw\_E[i]$ do
        if $j \notin neighbor\ of\ i$ then
            $E \mathrel{+}= (i, j)$
            $W_{i,j} \leftarrow metric(i, j)$
            $E \mathrel{+}= (j, i)$              ▷ undirected graph
            $W_{j,i} \leftarrow W_{i,j}$
        end
    end
end
```

Predicate == FALSE is better

4. TESTS

Distance metrics are compared in two major ways, in

- **Graph construction** using avg. # neighbors and the standard deviation.
- **Inference/classification** on the testing set for the LP_ZGL inference algorithm (Xiaojin et al. 2002).

Table 2 summarizes the test parameters.

Table 2 Test parameters.

Parameter	Value
# desired neighbors or nn	4
Testing/training ratio	10:1 to 40:1

Note that the high testing/training ratios were possible because of spectral transformation of the graph, as shown in Figure 5.

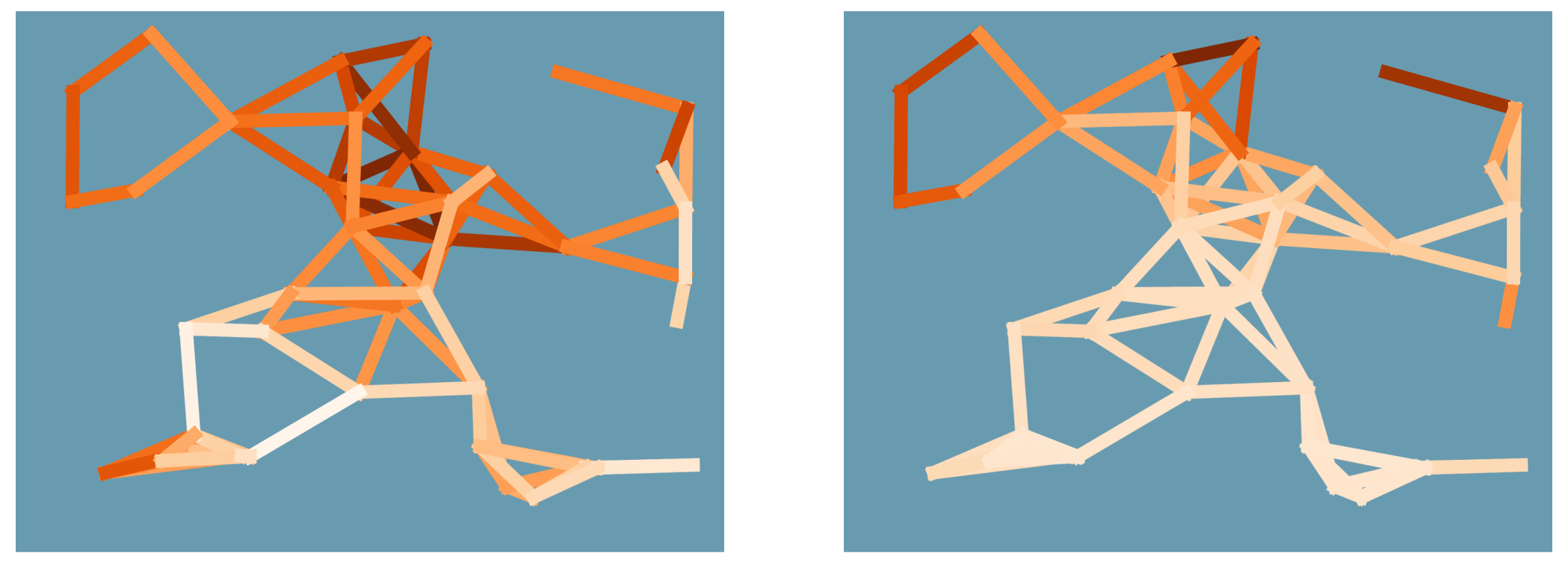


Figure 5 Spectral transformation using Gaussian field kernel. Original graph on the left ,and transformed on the right.

5. FINDINGS

Results are presented in Figure 6 and 7, and

- ISC (at around 1.5) gives closest to scaled $nn = 1.0$.
- Standard deviation is decisively minimum for ISC.
- ISC performs best in prediction accuracy.

Overall, performance trend of ISC > CS > L2 is evident.