



Enhancing Clustering Methods by Inserting Relational Information

Tijn Witsenburg¹

Hendrik Blockeel^{1,2}

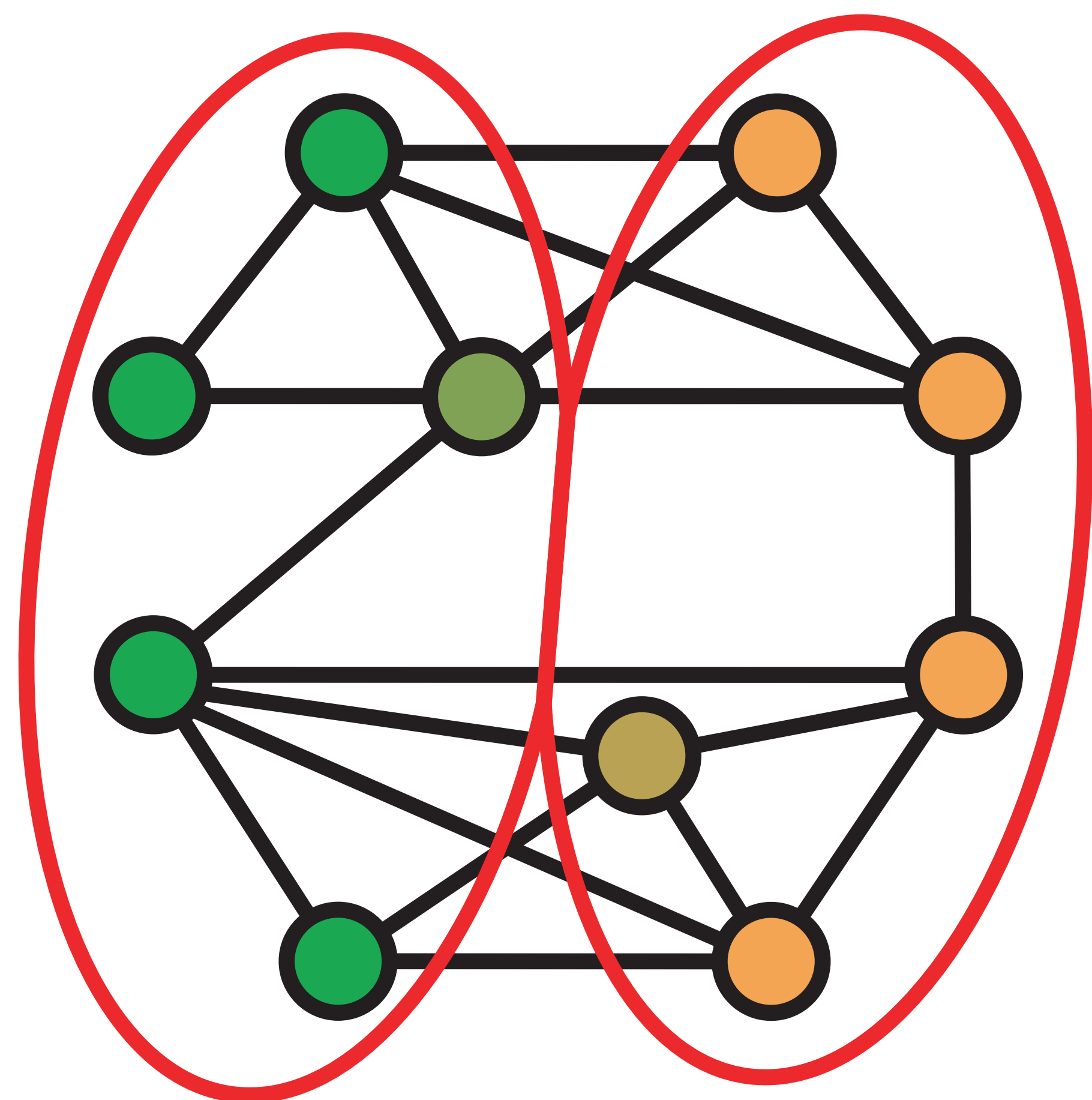
Universiteit Leiden

¹ Leiden Institute of Advanced Computer Science, Universiteit Leiden

² Department of Computer Science, Katholieke Universiteit Leuven

Main Goal

Can different types of information be combined to enhance clustering methods?

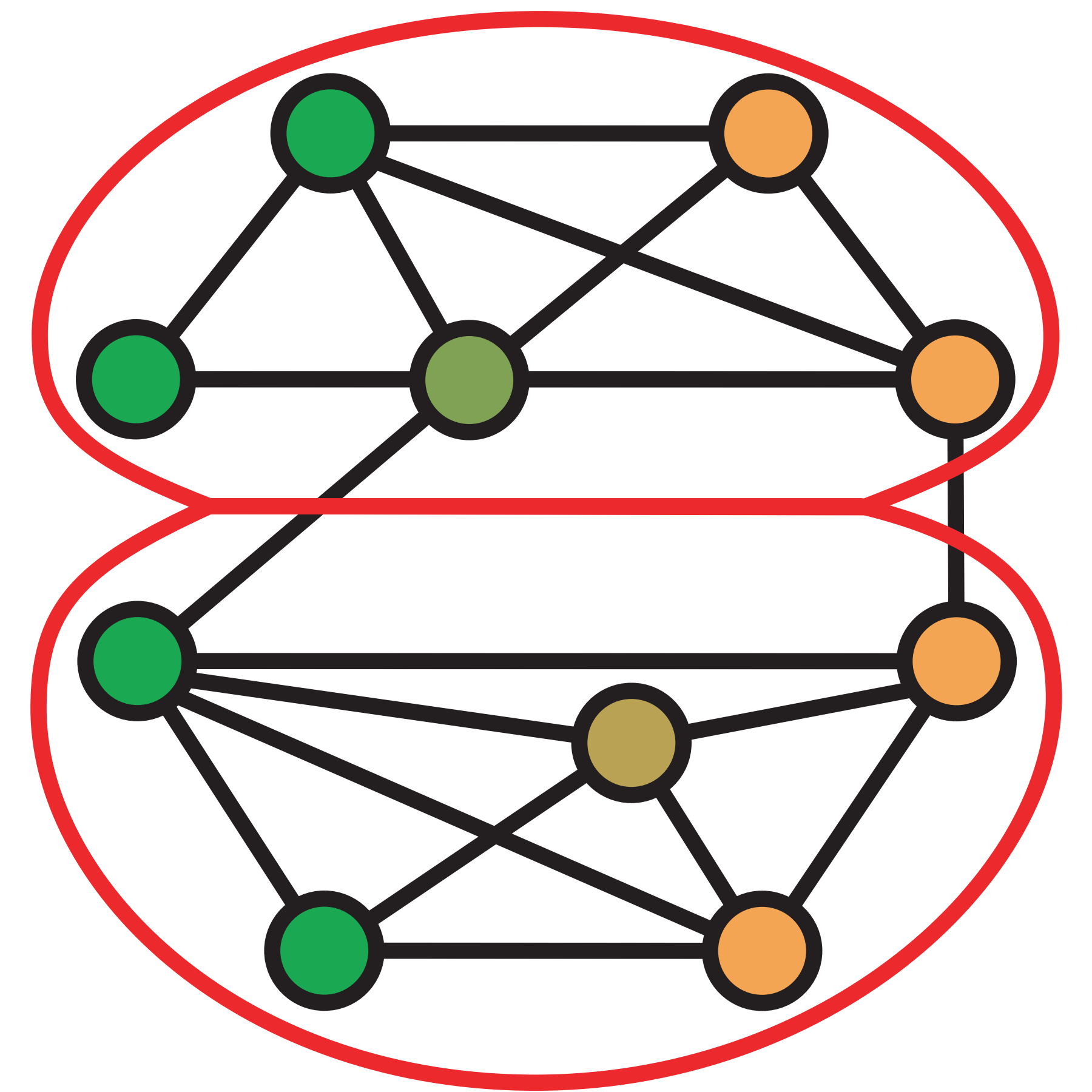


Different Types of Clustering Methods

Current clustering methods can be roughly divided into two types:

based on the content
(the colour of the node)

based on the relations
(the edges)



Cluster using a **Distance Matrix** where each element is the distance or similarity between two nodes.

Our Method

“Multiply” matrices for an average distance/similarity of the neighbours. Then use this new distance to alter the distance matrix and continue the clustering method normally.

Cluster using an **Adjacency Matrix** where each element is ‘1’ when a relation between two nodes exists.

Formal Aspects

Considering Similarity Matrix S and Adjacency Matrix A , then each element of new matrix M can be created using:

$$m_{ij} = c_1 \cdot s_{ij} + c_2 \cdot asn_{ij}$$

with c_1 and c_2 constants defining the ratio between the direct similarity (s_{ij}) and the average similarity of the neighbours (asn_{ij}) which has two versions:

$$asnI_{ij} = \frac{\sum_{k=1}^N s_{ik} \cdot a_{kj} + \sum_{k=1}^N s_{jk} \cdot a_{ki}}{\sum_{k=1}^N (a_{kj} + a_{ki})}$$

where the ratio between the amount of neighbours of i and j influences $asnI_{ij}$, and:

$$asnII_{ij} = \frac{1}{2} \cdot \left(\frac{\sum_{k=1}^N s_{ik} \cdot a_{kj}}{\sum_{k=1}^N a_{kj}} + \frac{\sum_{k=1}^N s_{jk} \cdot a_{ki}}{\sum_{k=1}^N a_{ki}} \right)$$

where the average similarity of the neighbours of i to j have the same influence as the other way around, despite their ratio.

First Results

First experiments on the Cora data set (scientific papers where abstracts are the content and citations the relations)

showed a significant improvement in cluster quality, while for experiments on IMDB (films where the plots are the content and having an actor in common is considered the relation) this was not yet so the case.

Questions:

- What could be the main reason for a **difference in results** between the two data sets?
- What should be the **constraints for the data** on which this method can be applied?
- How can **other information** be included as well?
- Can this principle be used on **other research fields**?
- Can people **help me find data sets** that have content, relations and classifications?

