Hyperlink Prediction in the Hypergraphs of Network Terms and Citations in Harrison C. White's 'Identity and Control'

Moses Boudourides

Master's in Data Science Online Program
School of Professional Studies
Northwestern University

Moses.Boudourides@northwestern.com

DGNet Congress

Netzwerkstrukturen zwischen Theorie und Praxis

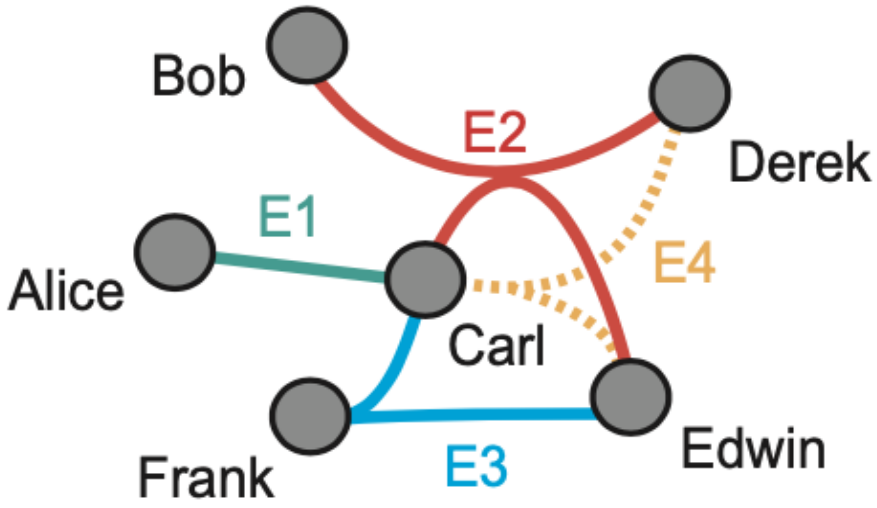
White–Panel 1

Monday, October 28, 2024

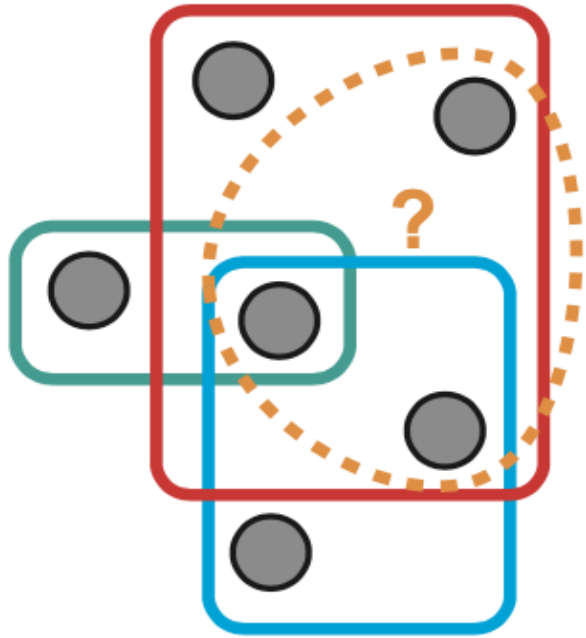
Hypergraphs

- A hypergraph is a generalization of a graph, in which the edges of a hypergraph, called hyperedges, can join any number of nodes (not only two as the edges of a graph can do).
- Formally, a **hypergraph** $H = (V, E)$ is a pairing of two finite sets, V and E, where V is the set of **nodes** (or **vertices**) and E, the set of **hyperedges** (or **hyperlinks**), is a set of (nonempty) subsets of V.
- **Node adjacency** is defined as the relation in V that a pair of nodes belong to the same hyperedge and **hyperedge adjacency** is defined as the relation in E that a pair of hyperedges (as sets of nodes) has nonempty intersection.
- If $|V| = n$ and $|E| = m$, the **incidence matrix** of the hypergraph is a $n \times m$ matrix of 1's and 0's, denoted as $\mathbf{H}$, such that if node v_i is involved in hyperedge e_j, then $\mathbf{H}_{ij} = 1$, while otherwise $\mathbf{H}_{ij} = 0$.
- The **degree of node** $i \in V$, denoted as d_i, is the number of hyperedges containg that node, i.e., $d_i = \sum_j \mathbf{H}_{ij}$, and the **cardinality of hyperedge** $e \in E$, denoted as c_j, is the number of nodes contained in that hyperedge, i.e., $c_j = \sum_i \mathbf{H}_{ij}$.

a



b



c

	E1	E2	E3	E4
Alice	1	0	0	0
Bob	0	1	0	0
Carl	1	1	1	1
Derek	0	1	0	1
Edwin	0	1	1	1
Frank	0	0	1	0

Hyperlink Prediction

- Hyperlink prediction in a hypergraph is the task of identifying missing hyperlinks that could plausibly exist within the hypergraph, based on the patterns in the observed structure and the attributes of the existing nodes.
- In general, hyperlink prediction is based on four categories of methods (see the review article of Chen & Liu, 2022 [https://doi.org/10.1109/TNNLS.2023.3286280]):
  - similarity-based,
  - probability-based,
  - matrix optimization-based, and
  - deep learning-based methods.

  Here, we are going to utilize one of the methods of the last category, called CHESHIRE, since neural networks significantly improve the performance of hyperlink prediction.

The CHESHIRE Method of Hyperlink Prediction (HP)

The **Chebyshev Spectral Hyperlink Predictor** (**CHESHIRE**) method frames hyperlink prediction as a machine learning classification problem over a set of hyperedges, including both *positive hyperedges*, i.e., existing hyperlinks observed in the hypergraph, and *negative hyperedges*, i.e., fake artificially generated (nonexistent) hyperlinks used for model balancing through **negative sampling**. CHESHIRE follows three main steps:

- Initializing node embeddings by simply passing the incidence matrix **H** through a one-layer neural network.
- Refining the node embeddings within hyperedges with a Chebyshev spectral graph convolutional network (GCN).
- Employing a Frobenius 2-norm-based pooling function to generate hyperlink embeddings.
- Finally, incorporating the maximum minimum-based pooling function as it is typically used by a neural hyperlink predictor (NHP).

In this way, CHESHIRE has effectively addressed the limitations of other deep learning-based methods (like HyperSAGCN and NHP) and achieved an outstanding performance on various types of hypergraph data (Chen & Liu, 2022; Chen, Liao & Liu, 2022 [`https://doi.org/10.1038/s41467-023-38110-7`]).

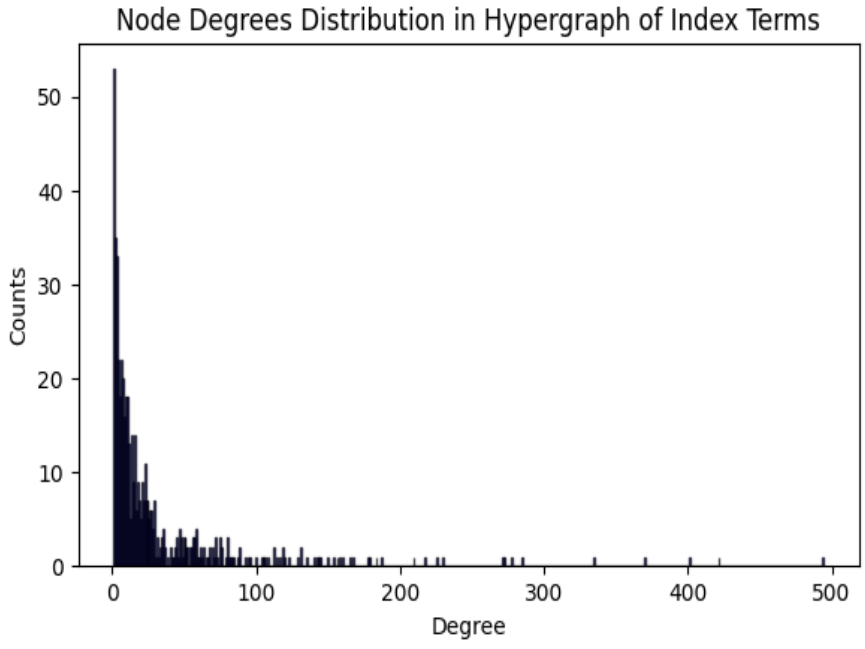
The Hypergraphs of White's 'Identity and Control' Book

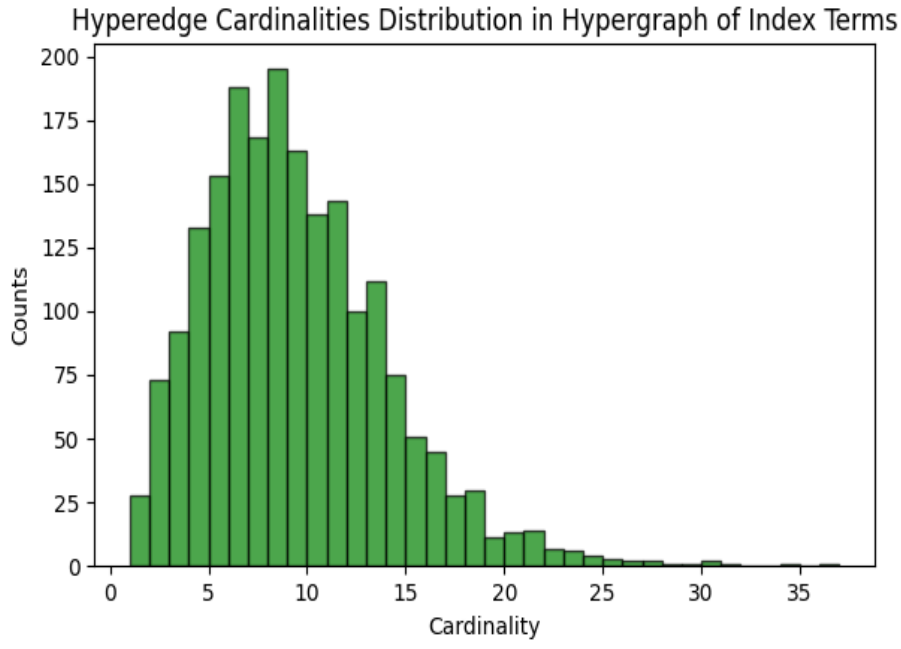
By segmenting the text of White's 'Identity and Control' into **paragraphs** and applying standard NLP techniques, we identified which **index terms** from the book's index and which **citations** from its references list co-occurred in each paragraph. Thus, two hypergraphs were constructed:

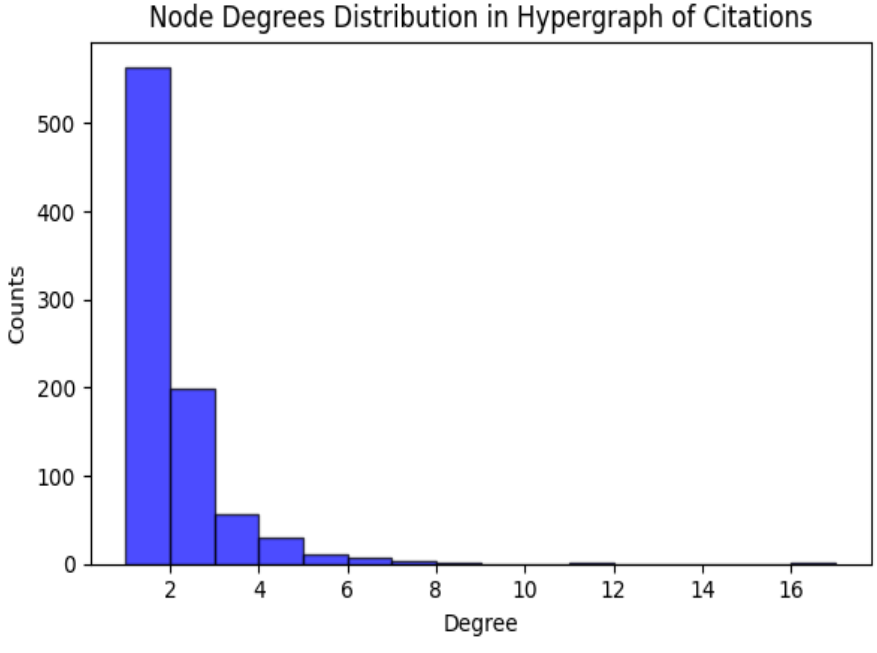
1. the **hypergraph of index terms**, where hyperedges represent paragraphs and nodes represent index terms, and
2. the **hypergraph of citations**, where hyperedges represent paragraphs and nodes represent citations.

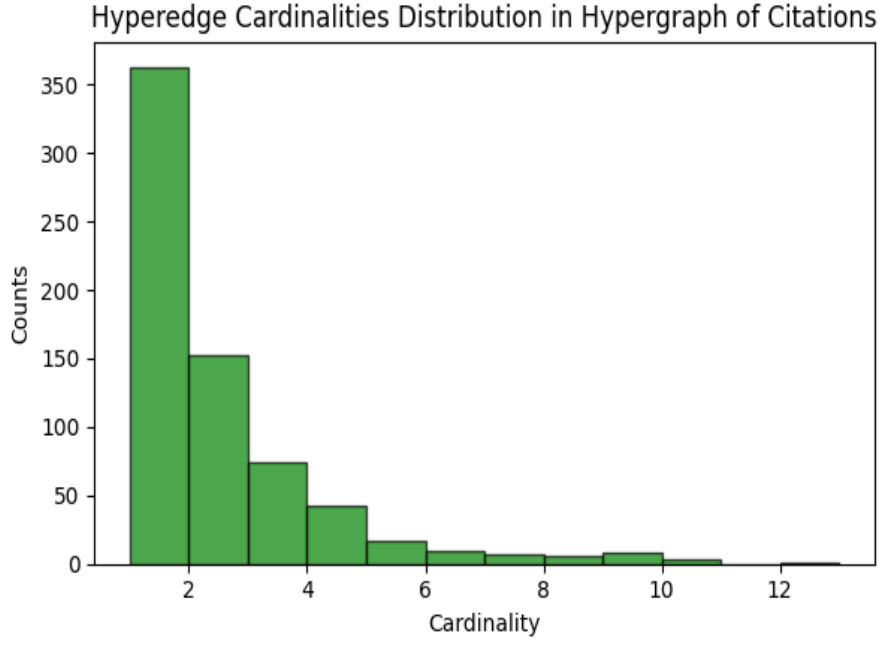
Hypergraph	# of nodes	# of hyperedges
Hypergraph of index terms	17788	1984
Hypergraph of citations	1409	683

The Histograms of Hypergraph Degrees and Cardinalities









The Feature of TF-IDF Scoring

Term Frequency-Inverse Document Frequency (`tf-idf`) is a numerical statistic used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It combines two measures: TF and IDF.

Components:

- **Term Frequency (TF):** Measures the frequency of a word in a document.

$$\text{TF} = \frac{\text{Number of times term } t \text{ appears in document}}{\text{Total number of terms in document}}$$

- **Inverse Document Frequency (IDF):** Reduces the weight of terms that appear in many documents, highlighting unique terms.

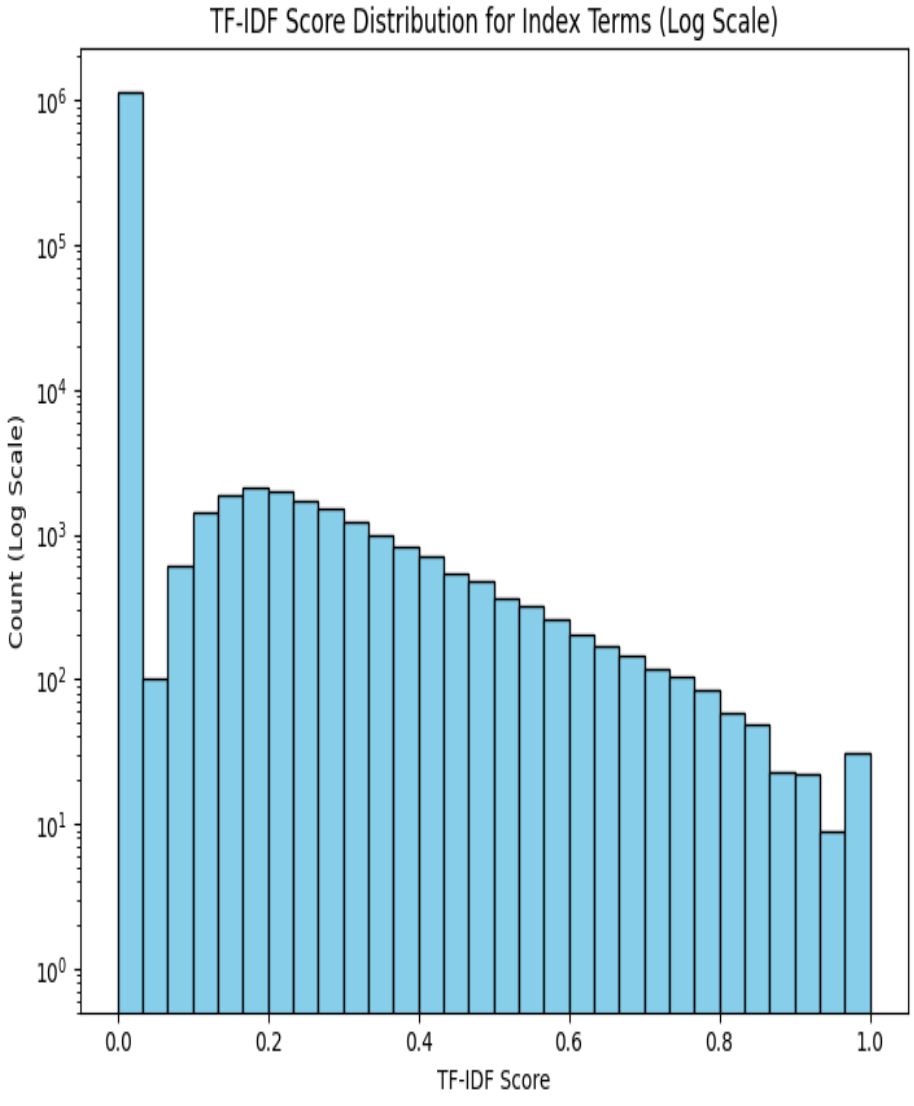
$$\texttt{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } t}\right)$$

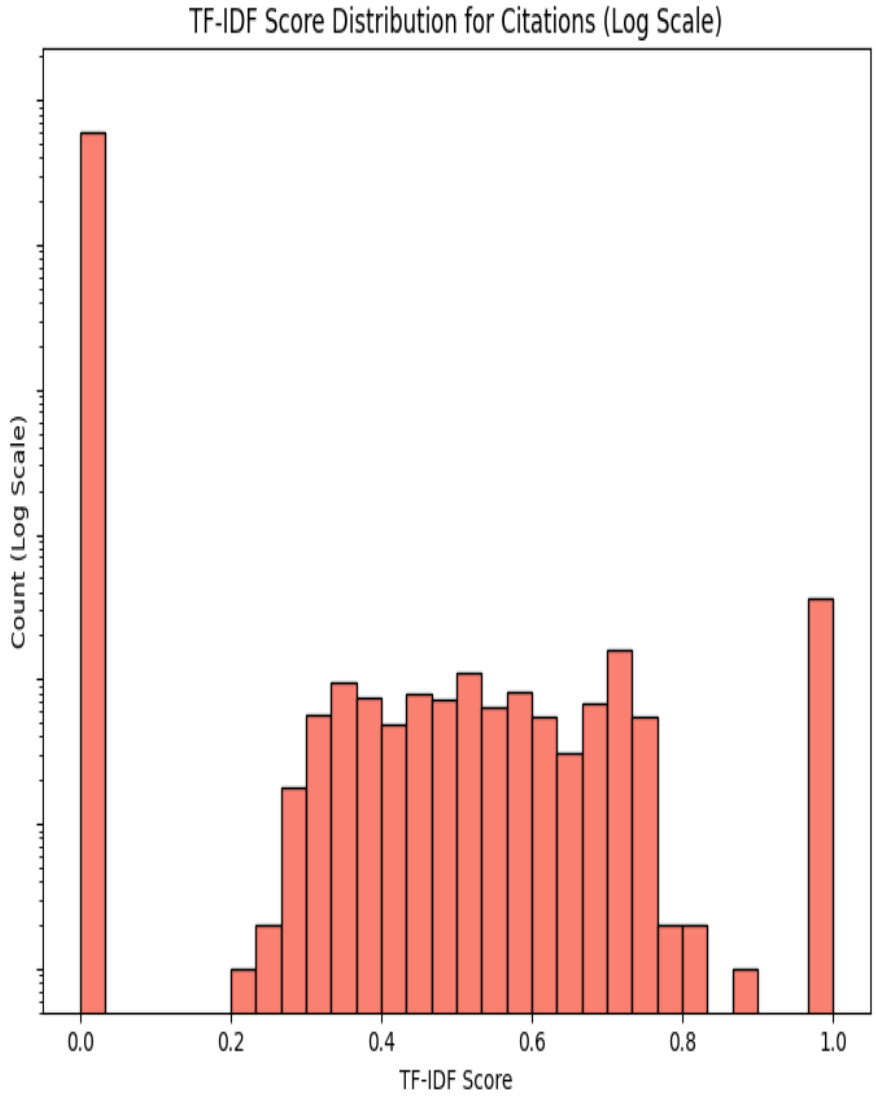
`tf-idf` Score: The score is calculated as the product of TF and IDF:

$$\texttt{tf-idf} = \text{TF} \times \text{IDF}$$

This score highlights terms that are important within a specific document but less common across the corpus.

The Histograms of the TF-IDF Features





CHESHIRE HP on White's 'Identity and Control' Book