A Flexible Weighting Framework For Converting Relational Database To Hypergraphs

Kotaro Fujii
Graduate School of Science and Technology
Keio University
Kanagawa, Japan
zico@inl.ics.keio.ac.jp

Tsuyoshi Yamashita
Graduate School of Science and Technology
Keio University
Kanagawa, Japan
tullys@inl.ics.keio.ac.jp

Andrew Shin
Faculty of Science and Technology
Keio University
Kanagawa, Japan
shin@inl.ics.keio.ac.jp

Kunitake Kaneko
Faculty of Science and Technology
Keio University
Kanagawa, Japan
kaneko@inl.ics.keio.ac.jp

***Abstract*—We propose a framework for converting relational databases (RDB) into hypergraphs to adjust and output various PageRank (PR) correlations by weighting hyperedges and nodes. While analyzing PR of converted graphs can reveal multifaceted information about relationships in the data, conventional methods struggled to balance the correlations between node PR and degree (NPR correlation) and edge PR and shared field records (EPR correlation). Our approach introduces two exponential weighting parameters: α for edge size, influencing EPR correlation, and β for node degree, influencing NPR correlation. By adjusting these parameters, various PR correlations can be obtained. Evaluation using real-world data demonstrate that our model can convert RDBs into hypergraphs with flexibility.**

***Index Terms*—Graph, Hypergraph, Database, PageRank**

I. Introduction

In recent years, there has been increasing interest in converting relational databases (RDBs) into graph databases for centrality index analysis, which can provide new insights into the relationships between data [12], [13]. Traditional RDB queries are limited in scope, as they focus on retrieving data directly rather than analyzing the connections between records. By transforming RDBs into graphs, it becomes possible to discover significant data relationships and perform secure analysis, as graphs allow data to be concealed by using node IDs rather than exposing full records. This approach enables users who do not have administrative access to analyze the data securely and gain insights into the database's structure.

PageRank (PR) centrality [10], which evaluates the importance of nodes and edges based on random walks (RW), is widely used in real-world applications such as social network analysis [11], recommendation systems [8], and community detection [5]. When RDBs are converted into graphs, PR can provide a valuable measure of the influence of specific records or fields. However, in existing methods, controlling the NPR correlation (node PR and degree correlation) and EPR correlation (edge PR and shared records correlation) has been challenging. Fields with a large number of records, such as those representing frequently appearing values, tend to dominate the PR analysis, inflating NPR and EPR values. As a result, top-ranked nodes and edges tend to be biased toward such fields, limiting the utility of the PR results.

To address this limitation, we propose a framework for transforming RDBs into hypergraphs, where weights are applied to nodes and edges to control NPR and EPR correlations. Hypergraphs, unlike conventional graphs, allow for the inclusion of multiple nodes in a single edge, and the weighting of these elements can be tailored to adjust the influence of specific fields on PR. By introducing exponential weighting parameters, α for edges and β for nodes, the proposed method adjusts the strength of these relationships, allowing for finer control over NPR and EPR correlations. For instance, to mitigate the impact of fields with many records, the weights for nodes and edges associated with these fields can be reduced, lowering their influence on PR without disregarding them entirely.

The framework was evaluated using real-world datasets, and the results showed that adjusting the α and β parameters allowed for the generation of weighted graphs with varying NPR and EPR correlations. Specifically, edge weighting primarily reduced EPR correlation, while node weighting contributed to the reduction of NPR correlation. This flexibility in controlling the PR correlations can produce more balanced and informative results, enabling deeper insights into the relational structure of the data. Moreover, applying other RW-based centrality measures, such as Random Walk Betweenness Centrality (RWBC) or Personalized PageRank (PPR) [4], [7], to the weighted hypergraphs could yield new perspectives, further expanding the analytical capabilities of graph-based RDB analysis.

II. Preliminaries

A hypergraph is a type of graph in which edges can connect more than two nodes. In addition to edge weighting, which

is commonly used in standard graphs, hypergraphs also allow for node weighting, enabling more advanced levels of analysis compared to traditional graphs. We refer to the number of nodes contained within an edge e as the edge size S_e.

Random walk on a hypergraph is executed in two phases [4]: the edge selection phase and the node selection phase. In edge selection phase, the random walker selects an edge containing the current node with a probability proportional to the edge weights. The probability P_{e_i} that edge e_i is selected is:

$$P_{e_i} = \frac{w_{e_i}}{\sum w_e} \tag{1}$$

where w_{e_i} is the weight of edge e_i, and $\sum w_e$ is the sum of the weights of all edges that include the current node. As shown in Equation 1, the probability of transitioning to a higher-weight edge is greater, while the probability of transitioning to a lower-weight edge is smaller.

In the node selection phase, a node contained within the selected edge is chosen with a probability proportional to the node weights. The probability P_{v_i} that node v_i is selected is:

$$P_{v_i} = \frac{w_{v_i}}{\sum w_v} \tag{2}$$

where w_{v_i} is the weight of node v_i, and $\sum w_v$ is the sum of the weights of all nodes contained within the selected edge. The probability of transitioning to a higher-weight node is greater, while the probability of transitioning to a lower-weight node is smaller. Finally, the random walker moves to the selected node, which becomes the new current node, and the edge selection process is repeated.

PageRank centrality is a metric that assigns higher values to nodes and edges that are more frequently reached by a random walker. The PR of a node is determined by the stay probability of the RWer at each node, while the PR of an edge is based on the stay probability at each edge. Unlike a standard RW, PR incorporates a termination probability, denoted as α. This termination prevents the RWer from getting trapped in loop structures within the graph, which would otherwise create biases in stay probabilities. In this paper, node PR is referred to as NPR, and edge PR as EPR.

III. Related Work

Converting relational databases (RDBs) into graphs offers two significant advantages: the ability to derive new insights and the possibility of secure data analysis. Traditional RDB queries focus on individual data points and lack the capability to uncover deep relationships between data. Graphs, on the other hand, excel at identifying important data and discovering strong connections between records, revealing value that cannot be achieved through conventional RDB queries. Additionally, graph analysis allows for secure handling of sensitive information by replacing data with IDs, enabling non-administrative users to perform analyses without requiring full data disclosure.

In methods where NPR (node PR and degree correlation) and EPR (edge PR and shared record correlation) are high, such as clique expansion (CE) [12], [13], records sharing the same field in an RDB are connected by edges, and RW transitions between connected nodes are equally probable. This leads to high NPR and EPR correlations, especially for fields with large numbers of records, resulting in nodes or edges from these fields dominating PR rankings. This method prioritizes degree, a single metric, limiting the ability to explore other features of the data, which can be problematic for users seeking more multifaceted insights.

On the other hand, hypergraph transformation [4] offers a method with low NPR and EPR correlations. Here, records are connected via hyperedges that correspond to field sizes, and transitions in RW are equally probable among edges and nodes. This method allows for the discovery of non-degree-based features, but the extremely low NPR and EPR correlations make it difficult for PR to reflect degree, an essential metric for graph analysis. As a result, users seeking degree-related insights may find this method unsatisfactory.

As such, existing methods for RDB-to-graph conversion exhibit polarized NPR and EPR correlations, limiting the diversity of information that PR can provide. We address this issue by proposing a method that adjusts NPR and EPR correlations through weighted transformations of nodes and edges. By introducing flexible weighting parameters, the proposed method enables PR to capture a wider range of correlations, offering more balanced insights compared to previous approaches.

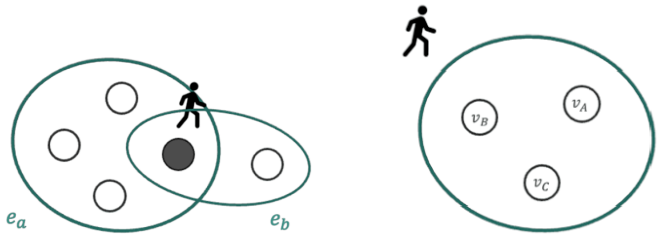
IV. Model

A. Overview

We now describe our proposed method, which extends the existing hypergraph transformation technique. The method converts the relational database (RDB) into a hypergraph and applies weighting to adjust the NPR and EPR correlations, thereby outputting a weighted graph with various PR correlation values.

Weighting Edges: The proposed method weights edges based on their size. To reduce both PR correlation values, the weights of larger edges are decreased, as these edges typically originate from fields with a high number of records. By altering the flow of the random walker, the influence of large edges on PR can be adjusted, allowing for the output of weighted graphs with varying PR correlations. Directly weighting larger edges will decrease the EPR correlation, and since nodes associated with larger edges tend to have high degrees, this also contributes to lowering the NPR correlation. This edge-focused weighting is expected to significantly impact the reduction of EPR correlation.

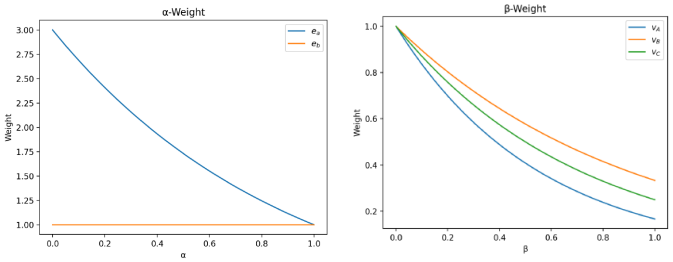
Weighting Nodes: Similarly, nodes are weighted according to their degree. To lower both PR correlation values, the weights of high-degree nodes are reduced. High-degree nodes often stem from fields with a large number of records, and adjusting the RWer's flow can modulate the influence of these nodes on PR, resulting in the generation of weighted graphs



(a) Example of a hypergraph with edge weighting

(b) Example of a hypergraph with node weighting

Fig. 1: Examples of hypergraphs to illustrate weighting of edges and nodes.



(a) Correlation between α and edges e_a,e_b

(b) Correlation between β and nodes v_A,v_B,v_C

Fig. 2: Correlation between each parameter and edges / nodes

with different PR correlation values. By directly applying weights to high-degree nodes, the NPR correlation is decreased. Additionally, since high-degree nodes are likely to be included in larger edges, this also contributes to reducing EPR correlation. This node-focused weighting is anticipated to particularly aid in lowering NPR correlation.

Given the user inputs for weighting parameters α and β for edges and nodes respectively, our model proceeds in the following manner: 1) The records of the RDB are transformed into nodes. Records that share the same field or fall within a specified numerical range in each column are connected by edges to create a hypergraph. 2) Edge and node weighting is applied based on the input parameters α and β. This process allows for the adjustment of NPR and EPR correlations, leading to a more nuanced analysis of the graph structure. 3) PageRank is computed on the transformed weighted hypergraph.

B. Weighting Nodes and Edges

1) Weighting Edges: We now describe in details how weights are assigned to edges. Since edge weights are determined based on the size of the edge, the weight w_{e_k} for an edge e_k is:

$$w_{e_k} = (S_{e_k} - 1)^{1-\alpha} \tag{3}$$

where S_{e_k} is the size of the edge, and α is the edge weighting parameter, with $0 \leq \alpha \leq 1$. Since the random walker here does not transition from the current node back to itself, the transition probability for an edge is proportional to $S_{e_k} - 1$. By Eq. 1, the probability P_{e_k} of transitioning to edge e_k is:

$$P_{e_k} = \frac{(S_{e_k} - 1)^{1-\alpha}}{\sum (S_{e_n} - 1)^{1-\alpha}} \tag{4}$$

where $\sum (S_{e_n} - 1)^{1-\alpha}$ is the sum of the weights of all edges containing the current node.

Figure 1a provides an example of a hypergraph with edge weighting. Let the current node be the black node, and assume that only edges e_a and e_b contain the current node. In this example, edge e_a has a size of 4, and edge e_b has a size of 2, leading to the following weights for edges e_a and e_b:

$$w_{e_a} = 3^{1-\alpha}, \quad w_{e_b} = 1^{1-\alpha} \tag{5}$$

Thus, the probability P_{e_a} of transitioning to edge e_a is:

$$P_{e_a} = \frac{3^{1-\alpha}}{3^{1-\alpha} + 1^{1-\alpha}} \tag{6}$$

Similarly, the probability P_{e_b} of transitioning to edge e_b is:

$$P_{e_b} = \frac{1^{1-\alpha}}{3^{1-\alpha} + 1^{1-\alpha}} \tag{7}$$

Figure 2a shows the relationship between the edge weighting parameter α and the weights of edges e_a and e_b. As the figure indicates, applying weights causes the relative weight of edge e_a (the larger edge) to decrease compared to that of edge e_b (the smaller edge), reducing the flow of the random walker toward the larger edge. This enables us to balance the flow between edges of different sizes. As the edge weighting parameter increases, the difference in flow between edges of different sizes becomes smaller, and the two PR correlation scores decrease. In particular, when $\alpha = 0$, the weights for edges e_a and e_b are:

$$w_{e_a} = 3, \quad w_{e_b} = 1 \tag{8}$$

Therefore, the transition probabilities P_{e_a} and P_{e_b} are:

$$P_{e_a} = \frac{3}{4}, \quad P_{e_b} = \frac{1}{4} \tag{9}$$

Since the random walker transitions proportionally to the edge size, it will transition to all adjacent nodes with equal probability. As a result, the random walker tends to transition to larger edges, resulting in the highest PR correlation.

When $\alpha = 1$, the weights for edges e_a and e_b are:

$$w_{e_a} = 1, \quad w_{e_b} = 1 \tag{10}$$

Thus, the transition probabilities P_{e_a} and P_{e_b} are:

$$P_{e_a} = \frac{1}{2}, \quad P_{e_b} = \frac{1}{2} \tag{11}$$

In this case, the walker transitions equally to all edges containing the current node, which reduces the probability of transitioning to nodes in larger edges. Thus, the PR correlation is minimized. By adjusting the edge weighting parameter between 0 and 1 based on user requirements, the proposed method can output random walks with various PR correlation values.

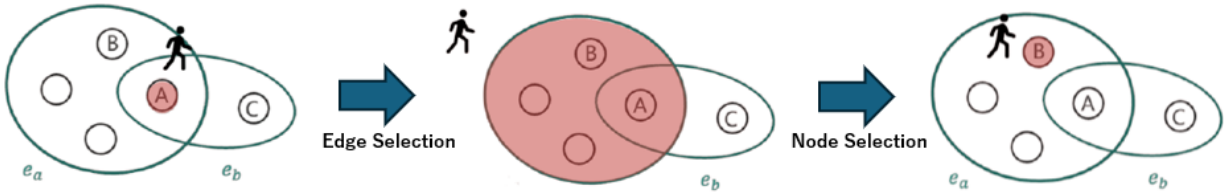


Fig. 3: Example of random walk transition in a weighted hypergraph.

2) Weighting Nodes: We now describe how weights are assigned to nodes. Since the weights are assigned to nodes based on their degrees, the weight w_{v_k} for a node v_k is:

$$w_{v_k} = d_{v_k}^{-\beta} \tag{12}$$

where d_{v_k} is the degree of node v_k, and β is the node weighting parameter, with $0 \leq \beta \leq 1$. From Eq. 2, the probability P_{v_k} of transitioning to node v_k is given by:

$$P_{v_k} = \frac{d_{v_k}^{-\beta}}{\sum d_{v_n}^{-\beta}} \tag{13}$$

where $\sum d_{v_n}^{-\beta}$ is the sum of the weights of all nodes contained in the current edge.

Figure 1b shows an example of a hypergraph with node weighting. Let the nodes v_A, v_B, v_C contained in the current edge have degrees of 6, 3, and 4, respectively. From Eq. 12, the weights of nodes v_A, v_B, v_C are:

$$w_{v_A} = 6^{-\beta}, \quad w_{v_B} = 3^{-\beta}, \quad w_{v_C} = 4^{-\beta} \tag{14}$$

Thus, the probability of transitioning from the current edge to node v_A, considering the weights of other adjacent nodes, is:

$$P_{v_A} = \frac{6^{-\beta}}{6^{-\beta} + 3^{-\beta} + 4^{-\beta}} \tag{15}$$

Similarly, the probabilities P_{v_B} and P_{v_C} of transitioning to nodes v_B and v_C are:

$$P_{v_B} = \frac{3^{-\beta}}{6^{-\beta} + 3^{-\beta} + 4^{-\beta}}, \quad P_{v_C} = \frac{4^{-\beta}}{6^{-\beta} + 3^{-\beta} + 4^{-\beta}} \tag{16}$$

Figure 2b illustrates the relationship between the node weighting parameter β and the weights of nodes v_A, v_B, v_C. As seen in the figure, applying weights causes the weight of the higher-degree node v_B to become relatively smaller compared to that of the lower-degree nodes v_A and v_C than before the weighting. This reduces the flow of the random walker to high-degree nodes compared to before weighting, thereby diminishing the difference in flow due to node degree. Moreover, increasing the node weighting parameter reduces the difference in flow between nodes of different degrees, leading to lower PR correlation between two nodes. In particular, when $\beta = 0$, the weights of nodes v_A, v_B, v_C are:

$$w_{v_A} = 1, \quad w_{v_B} = 1, \quad w_{v_C} = 1 \tag{17}$$

Thus, the probabilities $P_{v_A}, P_{v_B}, P_{v_C}$ of transitioning to nodes v_A, v_B, v_C are:

$$P_{v_A} = \frac{1}{3}, \quad P_{v_B} = \frac{1}{3}, \quad P_{v_C} = \frac{1}{3} \tag{18}$$

Since the transition probabilities to all nodes are equal, the random walk transitions to adjacent nodes with equal probability, resulting in the highest PR correlation.

On the other hand, when $\beta = 1$, the weights of nodes v_A, v_B, v_C become:

$$w_{v_A} = \frac{1}{6}, \quad w_{v_B} = \frac{1}{3}, \quad w_{v_C} = \frac{1}{4} \tag{19}$$

Thus, the probabilities $P_{v_A}, P_{v_B}, P_{v_C}$ of transitioning to nodes v_A, v_B, v_C are:

$$P_{v_A} = \frac{2}{9}, \quad P_{v_B} = \frac{4}{9}, \quad P_{v_C} = \frac{1}{3} \tag{20}$$

Since it becomes less likely for the random walker to transition to high-degree nodes, the PR correlation is minimized.

C. Transition of RW on Weighted Hypergraphs

In the hypergraph shown in Figure 3, let node A be the current location. The edge weighting parameter α is set to 0.4, and the node weighting parameter β is set to 0.3. Node A is assumed to belong only to edges e_a and e_b.

Consider the case of transitioning from node A to node B. In the random walk from node A to node B, the edge selection phase involves selecting edge e_a from the edges containing node A, and the node selection phase involves selecting node B from the nodes contained in edge e_a. The probability $P_{A \to e_a}$ of selecting edge e_a in the edge selection phase of the hypergraph RW is:

$$P_{A \to e_a} = \frac{3^{-0.6}}{3^{-0.6} + 1^{-0.6}} \tag{21}$$

The probability $P_{e_a \to B}$ of selecting node B in the node selection phase of the hypergraph RW is:

$$P_{e_a \to B} = \frac{6^{-0.3}}{6^{-0.3} + 8^{-0.3} + 4^{-0.3}} \tag{22}$$

The probability $P_{A \to B}$ of transitioning from node A to node B is the product of the probability $P_{A \to e_a}$ of selecting edge e_a in the edge selection phase and the probability $P_{e_a \to B}$ of selecting node B in the node selection phase. Therefore, from

Eq. 21 and Eq. 22, the probability $P_{A\to B}$ of transitioning from node A to node B is:

$$P_{A\to B} = \frac{3^{-0.6}}{3^{-0.6}+1^{-0.6}} \cdot \frac{6^{-0.3}}{6^{-0.3}+8^{-0.3}+4^{-0.3}} \tag{23}$$

V. Experiment

A. Setting

We now assess the effect of weighting both edges and nodes. Both α and β are varied between 0 and 1. We plot the changes in NPR correlation and EPR correlation when α and β are varied between 0 and 1, and evaluate whether combining the two parameters further reduces the two PR correlations. Additionally, by observing the NPR and EPR correlations for various combinations of α and β, we examine which parameter more strongly contributes to each PR correlation.

The evaluation method for PR correlations uses Spearman's rank correlation coefficient ρ [6]. Spearman's rank correlation coefficient ρ is a measure of the correlation between the ranks of two variables and is computed as:

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N} \tag{24}$$

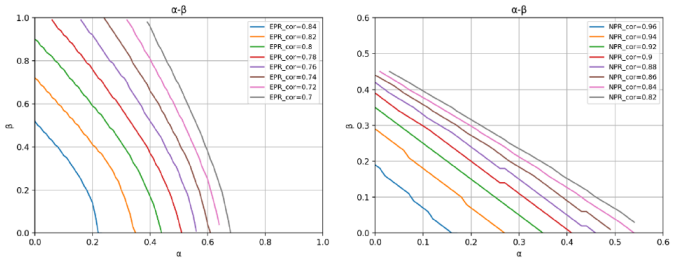
where D is the difference in ranks between the two variables, and N is the total number of nodes. NPR correlation is defined as the rank correlation coefficient between NPR and degree, and EPR correlation is defined as the rank correlation coefficient between EPR and edge size.

We evaluate the proposed method using three real-world graph datasets. Rakuten Recipes [1] is a relational database (RDB) that transforms recipe names into nodes and connects them with edges representing related tags and ingredients, consisting of 1,000 nodes and 1,563 edges. JAST Medical [3] is an RDB derived from prescription data where patients serve as nodes, while edges represent information related to the patients, such as diagnoses and age, making up 1,000 nodes and 1,071 edges. LIFFLE HOME [2] is an RDB constructed from property information where properties are represented as nodes, with edges connecting them to information such as municipalities, train stations, and structural features, adding up to 1,000 nodes and 1,125 edges.

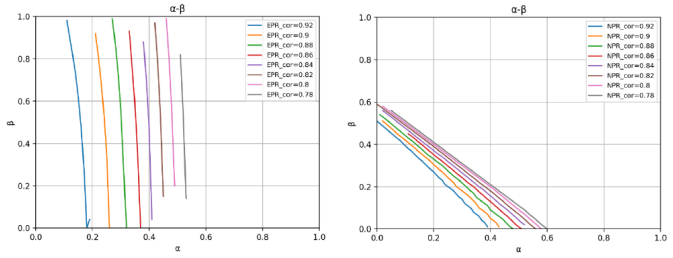
B. Results

Figures 4a, 4b, 4c show the relationship between α and β when the EPR correlation is fixed, and the relationship between α and β when the NPR correlation is fixed for each dataset. It is shown that for all datasets, combining edge weighting and node weighting further decreases the PR correlations. Additionally, it was found that increasing the values of the two weighting parameters results in a greater decrease in both PR correlations.

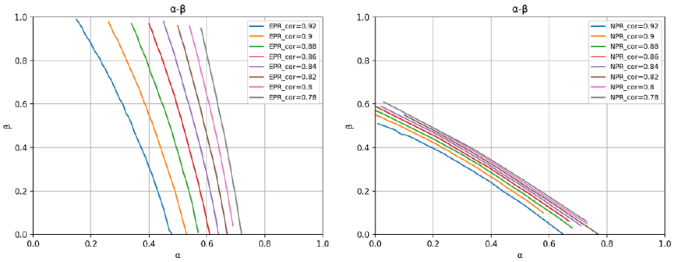
Figure 5a , 5b, 5c show the relationship between α and the EPR correlation, and between β and the EPR correlation when the NPR correlation is fixed, and the relationship between α and the NPR correlation, β and the NPR correlation when the EPR correlation is fixed. It is clear that in all datasets, when



(a) Correlation between α and β with EPR fixed (left), and with NPR fixed (right) for [1]



(b) Correlation between α and β with EPR fixed (left), and with NPR fixed (right) for [3]

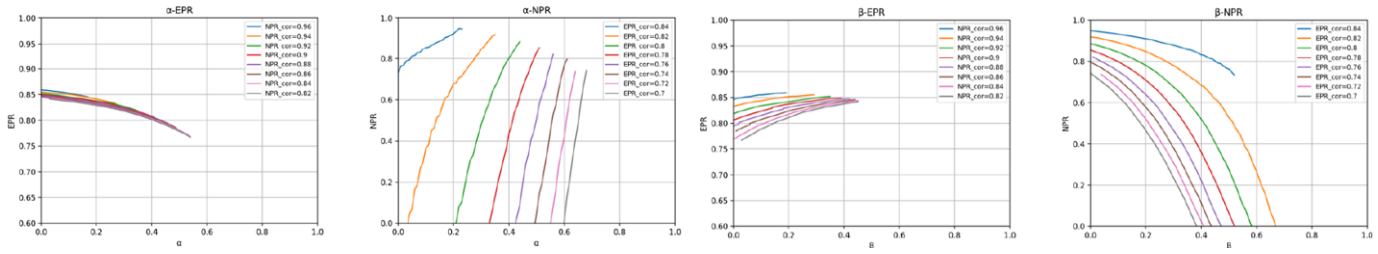


(c) Correlation between α and β with EPR fixed (left), and with NPR fixed (right) for [2]

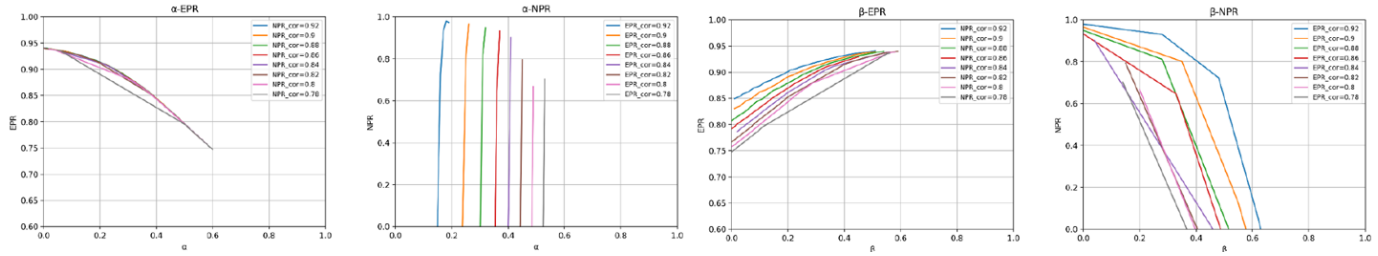
Fig. 4: Correlation between α and β with EPR / NPR fixed.

the NPR correlation is fixed, the EPR correlation decreases monotonically with respect to α, while it increases monotonically with respect to β. This indicates that α contributes more significantly to the decrease in EPR correlation compared to β. Furthermore, when EPR correlation is fixed across all datasets, the NPR correlation decreases monotonically with respect to β, while it increases monotonically with respect to α, indicating that β contributes more significantly to the decrease in NPR correlation compared to α.

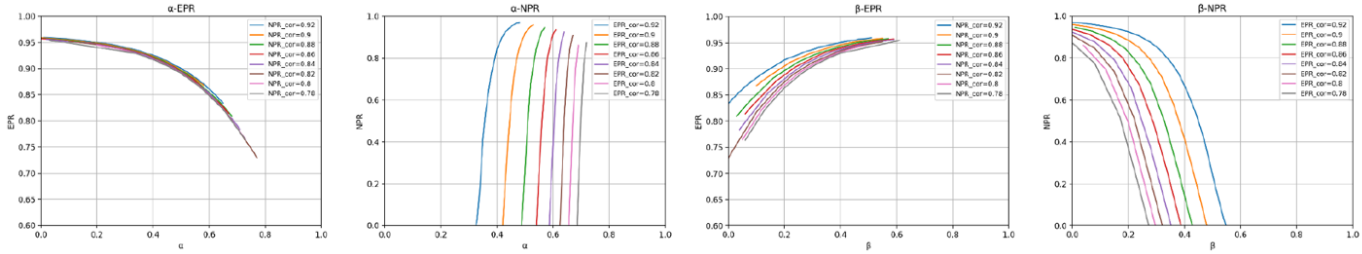
From the distribution of the weighting parameters when the PR correlation is fixed, it is clear that increasing the values of the weighting parameters α and β further decreases the two PR correlations. Furthermore, from the distribution of α and the EPR correlation when the NPR correlation is fixed, it is shown that α contributes more to the decrease in EPR correlation compared to β. Additionally, from the distribution of β and the NPR correlation when the EPR correlation is fixed, it is shown that β contributes more to the decrease in NPR correlation



(a) Correlation between the parameter and EPR / NPR for [1]



(b) Correlation between the parameter and EPR / NPR for [3]



(c) Correlation between the parameter and EPR / NPR for [2]

Fig. 5: Correlation between each parameter and EPR / NPR

compared to α.

VI. Conclusion & Future Work

This paper proposes a framework for converting relational databases into hypergraphs and applying edge and node weights to adjust NPR and EPR correlations, enabling the output of weighted graphs with varying PR correlations. Previous approaches suffer from lack of outcome controllability, leading to one-dimensional or less useful information. Our framework introduces weighting parameters based on edge size and node degree to fine-tune the correlations. Evaluation with real-world data showed that edge weighting primarily reduces EPR correlation, while node weighting reduces NPR correlation. New insights may be obtained by applying this approach to other metrics, such as Random Walk Betweenness Centrality [9] and Personalized PageRank (PPR) [7] for more diverse graph analysis.

References

[1] Rakuten Data Release. https://rit.rakuten.com/data_release/, 2012.
[2] Liffle home's dataset. https://www.nii.ac.jp/dsc/idr/lifull/, 2015.
[3] Jast medical data. https://www.jastlab.jast.jp/medical-data/, 2022.
[4] Uthsav Chitra and Benjamin J. Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *International Conference on Machine Learning*, 2019.
[5] Yang Gao, Xiangzhan Yu, and Hongli Zhang. Overlapping community detection by constrained personalized pagerank. *Expert Syst. Appl.*, 173:114682, 2021.
[6] Jan Hauke and Tomasz M. Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. 2011.
[7] Taher H. Haveliwala. Topic-sensitive pagerank. *Proceedings of the 11th international conference on World Wide Web*, 2002.
[8] Cataldo Musto, Pasquale Lops, Marco Degemmis, and Giovanni Semeraro. Context-aware graph-based recommendations exploiting personalized pagerank. *Knowl. Based Syst.*, 216:106806, 2021.
[9] Mark E. J. Newman. A measure of betweenness centrality based on random walks. *Soc. Networks*, 27:39–54, 2003.
[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*, 1999.
[11] Rui Wang, Weilai Zhang, Han Deng, Nanli Wang, Qing Miao, and Xinchao Zhao. Discover community leader in social network with pagerank. In *International Conference on Swarm Intelligence*, 2013.
[12] Konstantinos Xirogiannopoulos and Amol Deshpande. Extracting and analyzing hidden graphs from relational databases. *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017.
[13] Konstantinos Xirogiannopoulos, Udayan Khurana, and Amol Deshpande. Graphgen: Exploring interesting graphs in relational data. *Proc. VLDB Endow.*, 8:2032–2035, 2015.