# Constructing and Analyzing Different Density Graphs for Path Extrapolation in Wikipedia

Martha Sotiroudi, Anastasia-Sotiria Toufa, Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki, 54124, Greece
Email: {marthass, toufaanast, costas}@csd.auth.gr

***Abstract*—Graph-based models have become pivotal in understanding and predicting navigational patterns within complex networks. Building on graph-based models, the paper advances path extrapolation methods to efficiently predict Wikipedia navigation paths. The Wikipedia Central Macedonia (WCM) dataset is sourced from Wikipedia, with a spotlight on the Central Macedonia region, Greece, to initiate path generation. To build WCM, a crawling process is used that simulates human navigation through Wikipedia. Experimentation shows that an extension of the graph neural network GRETEL, which resorts to dual hypergraph transformation, performs better on a dense graph of WCM than on a sparse graph of WCM. Moreover, combining hypergraph features with features extracted from graph edges has proven to enhance the model's effectiveness. A superior model's performance is reported on the WCM dense graph than on the larger WIKISPEEDIA dataset, suggesting that size may not be as influential in predictive accuracy as the quality of connections and feature extraction. The paper fits the track Knowledge Discovery and Machine Learning of the 16th International Conference on Advances in Databases, Knowledge, and Data Applications.**

***Keywords*—*Wikipedia Dataset; Path Extrapolation; GRETEL; Dual Hypergraph Transformation; Graph Neural Networks.***

## I. INTRODUCTION

Graph structures offer an intuitive and powerful means to capture relationships and interactions within various kinds of data, paving the way for advanced analysis through the prism of Graph Neural Networks (GNNs) [1]–[5]. From node classification [6]–[8] to link prediction [9] [10], GNNs have proven indispensable across a spectrum of applications. Among these, the task of link prediction focuses on path inference, namely to predict an agent's trajectory over a graph.

The efficacy of such models is inherently tied to the quality and structure of the underlying graph. In this context, our work pivots on the creation of the Wikipedia Central Macedonia (WCM) dataset, a new dataset comprising paths extracted from the huge graph of Wikipedia, with a specific emphasis on articles related to Central Macedonia, Greece. The dataset tries to simulate human navigation paths as in WIKISPEEDIA [11] game, where users are asked to navigate from a given source to a given target article by only clicking Wikipedia links. Our objective is to leverage this dataset to address the problem of path inference.

WCM dataset is specifically designed to navigate through the complexities of Wikipedia's topology. It takes "Central Macedonia" as the starting article, from which it explores the external links through a series of random walks. Each step is contingent on a set of well-defined validity criteria. This ensures that each selected link is pertinent and non-redundant, providing a true reflection of the path an agent might traverse within the bounds of this thematic cluster. The dataset constructed for this study is made publicly available [12]. It comprises two separate files within the `Wikipedia_Dataset` directory, representing the `Dense Graph` and the `Sparse Graph` structures, each containing details of the paths, unique articles, path identifiers, categories, edges, hyperedges, observations, and path lengths. The code to create the WCM dataset can be found at [13].

The interest in the path inference problem has led to the development of advanced models like GRETEL [14], which has demonstrated promise in leveraging path extrapolation on graphs. GRETEL works as a generative model trying to capture the directionality of the path. It has been applied to both navigation data and paths constructed on the Wikipedia graph. This paper applies a graph transformation method based on the Dual Hypergraph Transformation (DHT) [15]. This method, as demonstrated in [16] [17], extends the traditional graph framework enabling connections among multiple nodes (i.e., vertices) within a hypergraph. Hypergraphs are suitable for this purpose because their edges can connect any number of nodes, not just two, as in a conventional graph. The new representation is able to capture more complex interactions between the data, and new more representative features can be extracted [18].

Here, in pursuit of advancing path extrapolation methods, WCM dense and sparse graphs are employed to assess both the original GRETEL and the Dual GRETEL variant in environments of varying complexity, providing a thorough insight into its adaptability and accuracy in different graph densities. To capture a comprehensive range of interactions within the data, a feature extraction process is implemented as proposed in [14] [16] [17]. [16] introduces an enhanced model, DualGRETEL+, that applies dual hypergraph transformation and a second-order optimizer to GPS navigation data, showing improved path inference capabilities. [17] assesses path extrapolation using GRETEL on Wikipedia data, with a focus on extracting informative features through the DHT.

The paper is structured as follows: Section II provides a detailed description of the dataset creation and its characteristics, along with an overview of the features employed and the

 ISBN: 978-1-68558-138-1 

GRETEL model. A detailed exposition of the experiments and results is found in Section III. The paper concludes in Section IV, underscoring the profound impact of graph density on the path extrapolation with graph neural networks.

## II. Methodology

This section focuses on the methodical approach to creating and analyzing the WCM, outlining the comprehensive process of collecting, categorizing, and extracting features from Wikipedia data to construct various graph types for path extrapolation.

### A. Dataset Creation

The dataset is created through a crawling process designed to traverse the vast interconnected landscape of Wikipedia, with Wikipedia Central Macedonia article [19] serving as the focal starting point. During data collection, we remained cognizant of the load implications on Wikipedia's servers. We inserted a pause of one second between two requests, safeguarding against potential server overload while accessing Wikipedia's data. This was a measure of digital courtesy and sustainability.

The path generation process begins with the Central Macedonia Wikipedia article. From this starting point, the crawler extracts all the external links associated with the current article. A subsequent article is then randomly selected from the set of external links, adhering to certain validity checks, ensuring the relevance of the link and its absence from the current path. To maintain the integrity of the dataset and concentrate solely on core articles, stringent validation criteria are instated. The process of path creation continues until the generated path either attains a predetermined length ranging from 4 to 7 articles or encounters an article devoid of valid external links. The algorithm employs a well-defined criterion to ensure the relevance and validity of each article within the path. The function `is_valid_title` is utilized to exclude titles containing terms like `Talk`, `User`, `File`, 'ISO', percentages, hashes, or colons, and those consisting solely of digits. This careful filtering is instrumental in maintaining a dataset focused on content-rich articles, avoiding disambiguation pages, meta-articles, or other forms of non-standard content that could detract from the dataset's integrity.

To ensure the intelligibility of the dataset, each Wikipedia article is associated with a distinct identifier. Leveraging tensor manipulation, the identifiers for the linked articles are distilled and organized within distinct tensor frameworks. These tensors serve as the foundation for the node indices within the constructed graph. To further aid our analysis, each trajectory's length is documented, and each article in the trajectory is associated with its unique identifier. This process is reiterated until a grand total of 3000 paths emerges. The graph $G$ is created comprising $m$ nodes and $n$ edges, where nodes represent articles and edges denote links between articles. The extracted paths are referred to as *trajectories*. We have documented these trajectories, noting their lengths and the articles they connect. The graph is represented using the Graph Markup Language.

Two distinct graph types have been created, each following a unique path selection process:

**Dense Graph**: This is formed by a modified path selection protocol within the crawler. Here, the crawler opts for a random choice from the first five external links of an article. We choose the first five external links for path selection to intentionally narrow down the possible trajectories, aiming for a denser graph structure that facilitates a more focused analysis of interconnected topics. This results in a connected network among a smaller subset of 912 nodes, 1311 edges, and 3000 paths. The same process of path generation, involving the extraction of links, applying validity checks, and documenting each trajectory with unique identifiers, is followed as in the general dataset creation.

**Sparse Graph**: This graph follows the initial broader selection process, incorporating a more extensive set of 7307 nodes, 10612 edges, and 3000 paths. The selection is made from all the external links.

### B. Article Categorization

Categorization provides a structured framework to analyze the dataset. Organizing articles into distinct categories enables researchers to identify content trends and patterns within the generated paths. This categorization not only enriches the dataset but also amplifies its potential utility for diverse research, analytical, and educational purposes.

Our categorization strategy focuses on dynamic online querying using DBpedia [20]. In order to determine the category of a given Wikipedia article, we rely on the SPARQL endpoint of DBpedia. Each article is queried to retrieve its semantic type from DBpedia's ontology. Whenever an explicit type is not obtained or if there are errors during the querying process, the articles are classified under `subject.General`.

### C. Feature Extraction

In addition to graph generation, a feature extraction process is conducted to leverage semantic information from the content of the articles and to capture complex interactions in the graph structure. According to [14], the feature vector for the nodes corresponds to its `in/out degree`, and its length is 2. For edges, the feature vector contains the Text Frequency - Inverse Document Frequency (`TF-IDF score`), capturing the semantic similarity between source and destination articles of a hyperlink [21], and the number of times the link was clicked in the training dataset of paths (`nof`).

*1) Dual Hypergraph Transformation*

The framework commences with the configuration of a conventional graph, designated as $G$ having $n$ nodes and $m$ edges. Node features are represented by a feature matrix $\boldsymbol{F} \in \mathbb{R}^{n \times d}$, and edge features by a feature matrix $\boldsymbol{E} \in \mathbb{R}^{m \times d'}$. Here, $d$ and $d'$ are the size of node and edge feature vectors, respectively. Considering an undirected graph, the incidence matrix is defined as $\boldsymbol{M} \in \{0, 1\}^{n \times m}$. In the case of a directed graph, the incidence matrix is defined as $\boldsymbol{M} \in \{-1, 0, 1\}^{n \times m}$.

In any case, the incidence matrix represents the relationships between nodes and edges in a graph, indicating which nodes are connected by specific edges.

The conventional graph and the corresponding dual hypergraph are represented as $G = (\boldsymbol{F}, \boldsymbol{M}, \boldsymbol{E})$ and $G^* = (\boldsymbol{F}^*, \boldsymbol{M}^*, \boldsymbol{E}^*)$ respectively. $\boldsymbol{F}^*$ represents the node features of hypergraph while $\boldsymbol{E}^*$ represents the hyperedge features. The DHT algorithm interchanges the roles of nodes and edges of the original graph [15]. That is, the edges of the original graph are reinterpreted as nodes in the dual hypergraph, while the original nodes become hyperedges in the dual hypergraph. Accordingly, $\boldsymbol{F}^* = \boldsymbol{E} \in \mathbb{R}^{m \times d'}$ and $\boldsymbol{E}^* = \boldsymbol{F} \in \mathbb{R}^{n \times d}$. The incidence matrix of the dual hypergraph is the transpose of the incidence matrix of the original graph, i.e., $\boldsymbol{M}^* = \boldsymbol{M}^\top$. The transformation is mathematically defined as:

$$G = (\boldsymbol{F}, \boldsymbol{M}, \boldsymbol{E}) \rightarrow G^* = (\boldsymbol{E}, \boldsymbol{M}^\top, \boldsymbol{F}) \tag{1}$$

Notably, the DHT is a reversible transformation, ensuring that applying it to $G^*$ recaptures the initial graph $G$, thereby preserving the structural and feature integrity of the transformation.

*2) Features extracted from the dual hypergraph*

Following the methodology proposed in [16], the original graph is transformed into its corresponding dual graph by applying the DHT algorithm in order to capture more complex interactions among edges. Two new features are extracted, namely the `similarity-hyperedge` and the `DHnode-in-out-degree`. The first feature assumes an undirected graph, while the second one assumes a directed graph. The implementation of dual hypergraph feature extraction, which significantly enhances the predictive accuracy of our models, can be found in [22].

For the `similarity-hyperedge` feature, the first step is to construct the incidence matrix $\boldsymbol{M} \in \{0,1\}^{n \times m}$. Row vector $\boldsymbol{q}_l \in \{0,1\}^m$ of $\boldsymbol{M}$, corresponds to node $l$. The cosine similarity between the incidence row vectors $\boldsymbol{q}_v$ and $\boldsymbol{q}_u$ is computed, where $v$ is the source node and $u$ is the target node of an arbitrary edge $e$. The corresponding vector in the $\boldsymbol{M}^*$ matrix is a column vector $\boldsymbol{q}_l^* \in \{0,1\}^m \equiv \boldsymbol{q}_l^\top$. The position of each 1 in this column vector indicates which nodes of the dual hypergraph are connected with the hyperedge $l^*$.

For the `DHnode-in-out-degree`, a directed graph $G$ is assumed. The corresponding incidence matrix is defined as $\boldsymbol{M} \in \{-1,0,1\}^{n \times m}$. To extract features associated with the input and output degrees of the dual hypergraph nodes, determining the direction of hypergraph edges becomes essential. This involves an examination of the column vector of $\boldsymbol{M}^*$ $\boldsymbol{q}_l^* \equiv \boldsymbol{q}_l^\top$. The position of each 1 in this column vector indicates which nodes of the dual hypergraph are connected with the hyperedge $l^*$. For every combination $(v_i^*, v_j^*)$, we verify the existence of a path $e_i \rightarrow e_j$ in the original graph that passes through the scrutinized node $l$. The new feature is the in-degree and out-degree of dual hypergraph nodes which are normalized by the maximum observed degree $D_{\max}$ in the hypergraph to facilitate comparison across different nodes:

$$\text{Normalized In/Out-Degree } \left(v_i^*\right) = \frac{\text{In/Out-Degree } (v_i^*)}{D_{\max}} \tag{2}$$

The aggregation of `similarity-hyperedge` and `DHnode-in-out-degree` results in `Similarity-Hyperedge-DHnode-In-Out-Degree` feature. These enhanced features are particularly critical in the sparse graph context, where the reduced number of connections demands a more nuanced approach to capturing node relationships. In the dense graph, with its inherently richer connectivity, these features play a pivotal role in distilling the essence of the network's complexity into a format conducive to advanced path prediction algorithms.

The feature extraction procedure is performed on the sparse graph with 7,307 nodes and 10,611 edges and the dense graph with 912 nodes and 1,311 edges.

### *D. Path Extrapolation Employing GRETEL*

The paper addresses path extrapolation focusing on predictive path analysis via the GRETEL model [14]. The graph $G$ consists of nodes and edges, represented as $G = (\mathcal{V}, \mathcal{E})$, with $n = |\mathcal{V}|$ denoting the node count and $m = |\mathcal{E}|$ the edge count, respectively. An agent progresses through the graph, stepping from node $v_i$ to $v_j$ contingent on the presence of a directed edge $e_{i \rightarrow j} \in \mathcal{E}$.

The agent's position at time $t$ is a sequential set of traversed nodes, symbolized as a given prefix $p = (v_1, v_2, \ldots, v_t)$. Let the path suffix $s = (v_{t+1}, \ldots, v_{t+h})$ be a collection of potential future for prediction horizon $h$. Within this setting, GRETEL is leveraged to estimate the conditional likelihood $\Pr(s \mid h, p, G)$ of path suffix $s$ given the prefix $p$, the horizon $h$, and the graph $G$. The agent's position at each step $t$ is encoded by a sparse vector $\mathbf{x}_t \in \mathbb{R}^n_{\geq 0}$ normalized to a unit sum, with its $i$-th element reflecting the likelihood of the agent being at node $v_i$.

GRETEL constructs a generative model that considers the directionality of edges via a latent graph with edge weights informed by a Multi-Layer Perception (MLP) that respects the graph's inherent directionality. The model's essence lies in its ability to forecast paths by learning from the traversed sequences, leveraging node features and the collective path history. More specifically, the non-normalized weights of each edge are computed by

$$z_{i \rightarrow j} = \mathrm{MLP}(\boldsymbol{c}_i, \boldsymbol{c}_j, \boldsymbol{f}_i, \boldsymbol{f}_j, \boldsymbol{f}_{i \rightarrow j}), \tag{3}$$

where $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are the pseudo-coordinates of the sender and the receiver node, respectively, while $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ denote the features of the sender and the receiver node, respectively. In (3), $\boldsymbol{f}_{i \rightarrow j}$ is the feature vector of the edge that connects the sender and the receiver node. The computed MLP outputs are normalized with the softmax function. The pseudo-coordinates $\boldsymbol{c}_i$ are computed using a GNN of $K$ layers. They are the agent representations $\boldsymbol{x}_\tau$ for $\tau \in \mathcal{I}$, where $\mathcal{I}$ denotes a trajectory. The non-zero elements of $\boldsymbol{x}_\tau$ refer to the distance between the agent and the $K$ closest graph nodes normalized to measure

one. Let $\vec{e}_t$ and $\vec{e}_t$ define the edges that go from $v_t \rightarrow v_{t+1}$ and $v_t \rightarrow v_{t-1}$, respectively. Let also $\boldsymbol{x}_t$ be the last position of the agent. GRETEL [14] can be trained through the *target likelihood*. That is, given a target distribution $\boldsymbol{x}_{t+h}$, the model tries to estimate the destination distribution $\hat{\boldsymbol{x}}_{t+h} \in \mathbb{R}^{n\times 1}$ over a horizon $h$ by the non-backtracking walk [23]

$$\hat{\boldsymbol{x}}_{t+h} = \boldsymbol{B}_\phi^+ \boldsymbol{P}_\phi^h \boldsymbol{B}_\phi \boldsymbol{x}_t. \tag{4}$$

Let $w_\phi(e_{k\rightarrow j})$ stand for the normalized MLP weights. In (4), $\boldsymbol{P}_\phi \in \mathbb{R}^{m\times m}$ has elements

$$[\boldsymbol{P}_\phi]_{e_{i\rightarrow j}, e_{k\rightarrow l}} = \begin{cases} 0 & \text{if } j \neq k \text{ or } i = l \\ \frac{w_\phi(e_{k\rightarrow l})}{1-w_\phi(e_{k\rightarrow i})} & \text{otherwise,} \end{cases} \tag{5}$$

$\boldsymbol{B}_\phi$ is a $m \times n$ matrix with $[\boldsymbol{B}_\phi]_{e_{i\rightarrow j},k} = 0$ if $k \neq i$ and $w_\phi(e_{k\rightarrow j})$, otherwise, and $\mathbf{B}_\phi^+$ stands for the pseudoinverse of $\mathbf{B}_\phi$. Such an approach integrates node and edge feature vectors, the former delineating the in/out-degree and the latter embedding the textual and usage-based similarity metrics. These primal features are pivotal in the model's capacity to estimate the suffix likelihood, aiding in approximating the path probability $\Pr(s \mid h, \varphi, G)$. In the paper, we will aggregate the original edge features $\boldsymbol{f}_{i\rightarrow j}$ with the features extracted from the dual hypergraph.

## III. Experiments and Results

To quantify the structure of each graph, we calculate the density, which provides a measure of how complete the graph is. The density is defined as the ratio of the number of edges $m$ to the number of possible edges, with the formula given for a directed graph without loops as

$$D = \frac{m}{n(n-1)}, \tag{6}$$

where $n$ is the number of nodes.

TABLE I. Dataset Characteristics

| Datasets | Nodes | Edges | Density |
|---|---|---|---|
| **Sparse Graph** | 7307 | 10612 | $2 \times 10^{-4}$ |
| **Dense Graph** | 912 | 1311 | $1.58 \times 10^{-3}$ |
| **Wikispeedia** | 4604 | 119882 | $5.66 \times 10^{-3}$ |

Table I summarizes the characteristics of the graphs used in the experiments, providing a clear comparison of the number of nodes, edges, and density across the Sparse Graph, the Dense Graph, and WIKISPEEDIA. Based on the characteristics outlined in Table I, the sparse graph demonstrates a lower density ratio due to its larger node count. In contrast, the dense graph, with fewer nodes, exhibits a higher density ratio. Notably, the WIKISPEEDIA dataset possesses the greatest density ratio of the three.

The following metrics are used to assess the feature vectors. *Target probability* measures the average chance that the model will choose a node with non-zero likelihood. *Choice accuracy* measures how accurate the decisions of an algorithm are at each crossroad of the ground-truth path, connecting nodes $v_t$ and $v_{t+h}$. It is computed on nodes whose degree is at least 3. *precision top1* measures how often the correct next step appears in the model's first prediction only, while *precision top5* evaluates how often the correct next step appears within the model's first five predictions.

In all experiments, the node feature vector includes the in/out degree for the nodes, retaining a constant size of two, underscoring consistent complexity in nodal characteristics despite the variation in graph densities.

An empirical assessment of model performance using the features derived from the original graph and those of the corresponding dual hypergraph is conducted. In the case of `original edges`, the `TF-IDF score` and `nof` features are used, yielding a feature vector of size 2. By aggregating the features `similarity-hyperedge` of length 1, `DHnode-in-out-degree` of length 2, and their combination `similarity-hyperedge-DHnode-in-out-degree` of length 3, the associated edge feature vector has length 3, 4, and 5, respectively.

Table II summarizes the performance of the GRETEL model with original edge features and the features extracted from the dual hypergraph (Dual GRETEL) added on top of the original edge features on the Sparse Graph.

Table III repeats the model's performance assessment on the Dense Graph. Table IV details the model's performance on the WIKISPEEDIA dataset. This dataset encapsulates the essence of human navigational strategies within Wikipedia, compiling 51318 completed paths from the WIKI GAME where participants navigate through article links towards a target article, with an aim for efficiency in both clicks and time.

The modularity class algorithm in `Gephi` [24] is used to identify the clusters within the network. These clusters contain nodes that are more densely connected to each other than to nodes in different clusters. The resulting clusters are indicated by the color coding of the nodes. The size of each node is proportional to its degree, reflecting the number of connections it has within the network. This allows for the immediate visual identification of highly connected nodes. The visible labels on the nodes in the figures were chosen because they have higher degree values, which show their importance in the graph, and they represent the main topic of each cluster within the expansive Wikipedia network.

Figure 1 represents the dense graph of Wikipedia. The selective navigation results in a dense network with several clusters, one of which is built around the Central Macedonia article, connecting closely related topics. Adjacent nodes like 'History of Greece' and 'Politics of Greece' form clusters that delve into the nation's past and governance, and 'Geographic Coordinate System' and 'France' appear as nodes indicative of broader geographical discourse.

The visualization of the sparse graph in Figure 2 reveals a network that unfolds from the Central Macedonia article, forming a large, primary cluster due to the random link selection strategy, and extending outward into a sparse array of smaller clusters. These smaller clusters are thematic, with subjects such as European countries, Greek cities, and historical events.

 ISBN: 978-1-68558-138-1 

TABLE II. PERFORMANCE METRICS (%) ON THE SPARSE GRAPH

| | GRETEL | Dual GRETEL | | |
|---|---|---|---|---|
| **Metrics** | **Original Edges** | **Similarity-Hyperedge** | **DHnode-In-Out-Degree** | **Similarity-Hyperedge-DHnode-In-Out-Degree** |
| **target probability** | $68.76 \pm 0.0044$ | $68.76 \pm 0.0019$ | $68.99 \pm 0.0064$ | $\mathbf{69.71 \pm 0.0038}$ |
| **choice accuracy** | $\mathbf{51.18 \pm 0.0011}$ | $38.69 \pm 0.0042$ | $39.60 \pm 0.0082$ | $39.24 \pm 0.0090$ |
| **precision top1** | $66.65 \pm 0.0050$ | $66.71 \pm 0.0012$ | $66.65 \pm 0.0025$ | $\mathbf{67.14 \pm 0.0045}$ |
| **precision top5** | $80.62 \pm 0.0019$ | $80.62 \pm 0.0019$ | $80.68 \pm 0.0023$ | $\mathbf{80.98 \pm 0.0036}$ |

TABLE III. PERFORMANCE METRICS (%) ON THE DENSE GRAPH

| | GRETEL | Dual GRETEL | | |
|---|---|---|---|---|
| **Metrics** | **Original Edges** | **Similarity-Hyperedge** | **DHnode-In-Out-Degree** | **Similarity-Hyperedge-DHnode-In-Out-Degree** |
| **target probability** | $0.0030 \pm 0.0021$ | $\mathbf{19.1007 \pm 0.0004}$ | $18.8741 \pm 0.0033$ | $19.0980 \pm 0.0026$ |
| **choice accuracy** | $\mathbf{48.0602 \pm 0.0135}$ | $27.8261 \pm 0.0084$ | $29.8662 \pm 0.0096$ | $29.5318 \pm 0.0086$ |
| **precision top1** | $0.001 \pm 0.0023$ | $19.8995 \pm 0.0074$ | $18.0904 \pm 0.0075$ | $\mathbf{20.5025 \pm 0.0067}$ |
| **precision top5** | $0.2513 \pm 0.0012$ | $83.2161 \pm 0.0088$ | $82.8141 \pm 0.0258$ | $\mathbf{83.8694 \pm 0.0112}$ |

TABLE IV. PERFORMANCE METRICS (%) ON THE WIKISPEEDIA DATASET

| | GRETEL | Dual GRETEL | | |
|---|---|---|---|---|
| **Metrics** | **Original Edges** | **Similarity-Hyperedge** | **DHnode-In-Out-Degree** | **Similarity-Hyperedge-DHnode-In-Out-Degree** |
| **target probability** | $6.42 \pm 0.1$ | $\mathbf{6.74 \pm 0.1}$ | $6.44 \pm 0.2$ | $6.2 \pm 0.1$ |
| **choice accuracy** | $22.16 \pm 0.4$ | $\mathbf{23.2 \pm 0.1}$ | $22.88 \pm 0.1$ | $21.86 \pm 0.4$ |
| **precision top1** | $11.6 \pm 0.2$ | $\mathbf{12.7 \pm 0.1}$ | $12.14 \pm 0.1$ | $11.66 \pm 0.3$ |
| **precision top5** | $30.1 \pm 0.1$ | $\mathbf{30.14 \pm 0.1}$ | $30.02 \pm 0.05$ | $30 \pm 0.09$ |

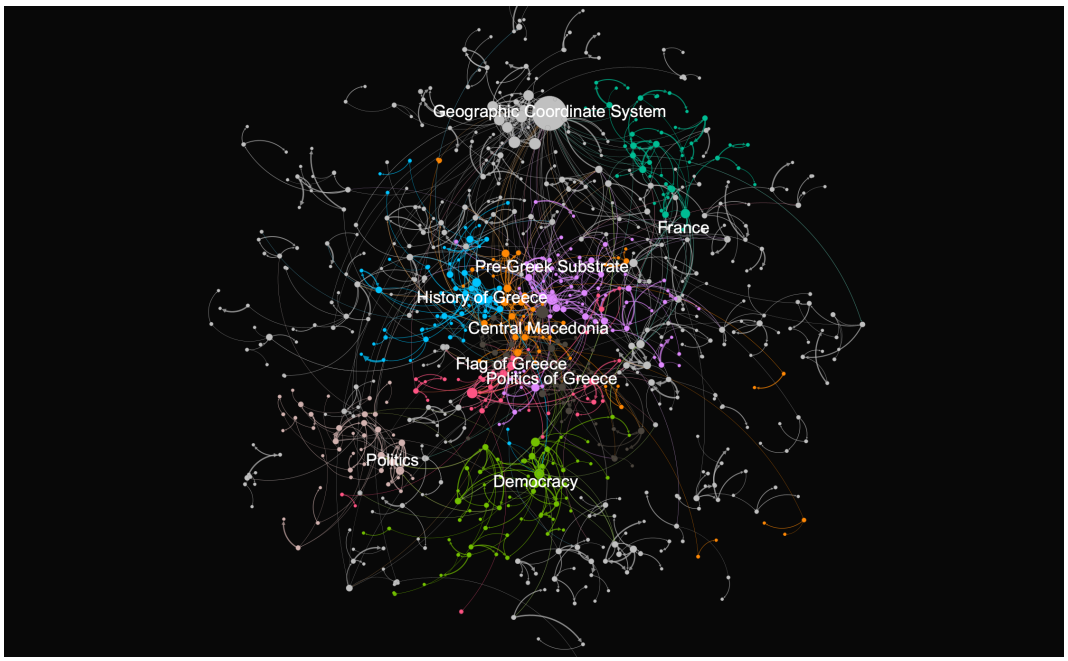


Figure 1. Dense Wikipedia Graph.

The construction methods of the two graphs distinctly shape their representations. The dense graph demonstrates that the Central Macedonia article forms a cluster, with surrounding clusters closely related in theme, predominantly focusing on Greece. This clustering suggests that the used method tends to group related topics tightly together. On the other hand, the sparse graph shows a different pattern where the Central Macedonia article and closely linked articles stand out in num-

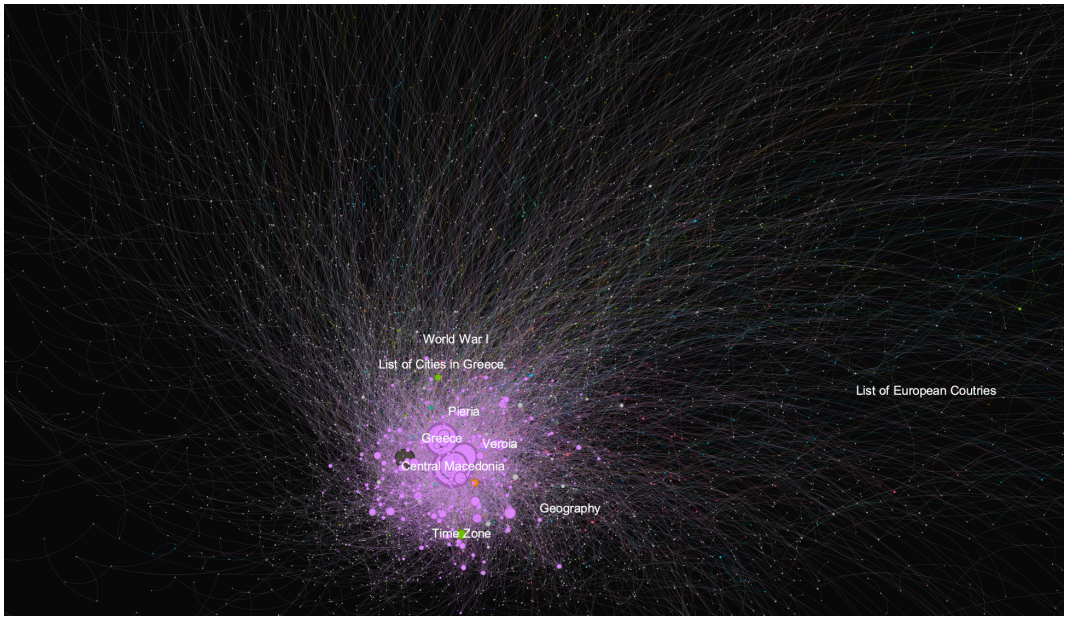


Figure 2. Sparse Wikipedia Graph.

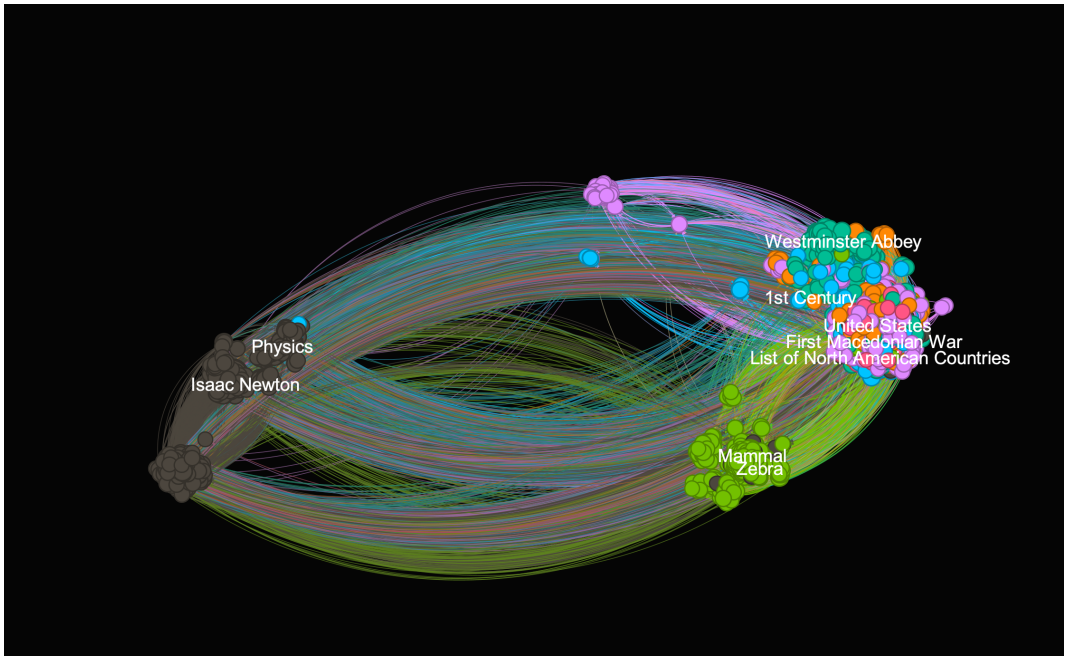


Figure 3. Wikispeedia Graph.

ber, while other articles appear less connected. This difference highlights how the choice of links in the construction process can significantly affect the network's structure.

Figure 3 represents the Wikispeedia graph, characterized by uniformly sized nodes, indicative of a network without a predominant starting article. Clusters within the graph are thematically organized, with 'Isaac Newton' and 'Physics' forming a cluster around scientific inquiry, while 'Westminster Abbey' serves as a node for the cluster concerning England. 'Mammal' and 'Zebra' are central to a cluster on zoology. These labels serve as the focal points for their respective clusters, marking the diverse subjects navigated by users.

Table V showcases examples of how the extracted features are employed to predict specific paths, highlighting the model's ability to deduce the most probable outcomes. Table V includes the conditional probabilities that reflect the model's ability to correctly anticipate the actual path taken. These examples are instrumental in illustrating the practical application of the model and the effectiveness of the features in guiding the model toward the most probable navigational

 ISBN: 978-1-68558-138-1 

route. The utilization of hypergraph features results in higher conditional probability compared to the use of `original edge` features. The examples clearly show that when hypergraph features are considered, the model tends to assign a greater likelihood to the true path, suggesting that these features capture more of the complexities inherent in human navigational behaviors on Wikipedia. The examples are drawn from the sparse graph of the WCM dataset.

### A. Performance Analysis of Model Across Dense and Sparse Graphs

In the evaluation of the Dual GRETEL model, distinct performances are observed between the sparse and dense graphs. A higher predictive accuracy with respect to *precision top5* metric is measured for the dense graph than the sparse one. This improved performance can be attributed to the vital role of the hyperedges, which enrich the model's contextual framework for more accurate extrapolation.

The sparse graph, despite its lower connectivity, shows commendable results, outperforming the dense graph in terms of target probability and choice accuracy. Dual GRETEL predicts the correct target with a probability of $69.71 \pm 0.0038$ %. GRETEL accurately chooses the next step with a rate of $51.18 \pm 0.0011$ %. It's noteworthy that except for the precision top 5, Dual GRETEL maintains a better performance on the sparse graph than the dense one.

This comparison reveals that while the Dual GRETEL model benefits from the rich link structures in dense graphs for precision tasks, it retains substantial predictive strength in sparse settings. This insight may guide further optimization for the model, enhancing its adaptability across varying network densities.

Also, this performance indicates that the model may benefit from the reduced complexity in sparse networks, potentially due to less noise and fewer connections, which can simplify the path prediction process. The comparison suggests that the model might generalize better in sparse environments, avoiding potential overfitting that can occur in dense networks with more intricate connections. Conversely, the specificity that dense networks provide can enhance the model's precision in certain contexts.

### B. Model Benchmarking on WCM Dense Graph Versus Wikispeedia Graph

For the WCM dense graph, Dual GRETEL demonstrated a significant improvement, achieving an impressive precision top5 score of $83.8694 \pm 0.0112\%$. On the WIKISPEEDIA graph, Dual GRETEL also showed enhanced performance with a precision top5 score of $30.14 \pm 0.1\%$. This indicates that hypergraph features greatly enhance the model's ability to accurately identify the most likely paths in a dense environment.

Furthermore, GRETEL demonstrated a high choice accuracy of $48.0602 \pm 0.0135\%$ on the dense graph compared to $23.2 \pm 0.1\%$ of Dual GRETEL on the WIKISPEEDIA dataset. Our findings show that model performance on the dense graph improves across all metrics except *choice accuracy* when we use hypergraph features. That is, hypergraph features are particularly effective in densely connected graphs, enhancing the model's predictive accuracy across all metrics we tested. The results indicate the potential of hypergraph features to improve the performance of path prediction models like GRETEL, especially in complex network structures.

The completion rate of paths in the WIKISPEEDIA dataset may introduce additional complexity, given that there is a mixture of successful and abandoned paths. In contrast, the smaller dataset might offer more uniformly successful paths, influencing the ease with which the model can learn and predict.

The analysis of the model's performance, as shown in Tables II-IV, reveals a trend where effectiveness inversely correlates with graph density. This suggests that as graphs become more interconnected, the model encounters greater challenges in path prediction accuracy. These observations emphasize the critical role that graph density plays in the deployment and refinement of path prediction algorithms. A possible explanation for the deterioration of accuracy as density increases could be the rise in potential paths that the model must discern. In denser graphs, the increased interconnectivity results in a greater number of plausible trajectories between nodes, potentially complicating the model's task of pinpointing the most likely path. Furthermore, a dense network may introduce more noise in the form of less relevant or weaker connections, which could mislead the prediction algorithm. These findings indicate that models like GRETEL or Dual GRETEL may require adjustments or enhancements, such as more sophisticated feature extraction or the incorporation of context-aware learning mechanisms, to better handle the complexity introduced by higher-density graphs.

## IV. CONCLUSIONS

A detailed analysis of GRETEL and its variant Dual GRETEL has been presented on dense and sparse graphs derived from the WCM dataset, aiming to improve path extrapolation models. Having developed the novel dataset centered on Central Macedonia, Greece, we have provided a resource that captures the complexity of human navigational patterns on Wikipedia.

Our investigation has shown that the density of a graph significantly influences the effectiveness of path prediction methods. Both models have performed better on sparse graphs in various aspects, yet they have achieved higher accuracy with respect to the top five predictions on the dense WCM graph. Furthermore, the incorporation of hypergraph features into the GRETEL model yielding the Dual GRETEL variant has significantly enhanced the accuracy of path predictions, underscoring the importance of feature extraction in graph-based predictive analytics. Comparisons of Dual GRETEL performance on the more extensive WIKISPEEDIA dataset against the WCM dense graph have also shown that the top metrics were measured on the WCM dense graph, despite its smaller size. This indicates that the model's success is

 ISBN: 978-1-68558-138-1 

TABLE V. EXAMPLES OF PATH PREDICTION

| | | $\Pr(s \mid h, p, G)$ | | $\Pr(s \mid h, p, G)$ | | $\Pr(s \mid h, p, G)$ |
|---|---|---|---|---|---|---|
| **prefix** | **Naousa, Imathia,<br>History of Macedonia, Craterus** | | **Volvi, Egnatia,<br>Thessaloniki, Arethousa** | | **Thessaloniki, Greek National Road,<br>Evzonoi, Axioupoli** | |
| **true suffix** | **Antigenes, Nearchus, Tlepolemus** | | **Nea Madytos, Vrasna** | | **Greek Macedonia, Despotate of Epirus** | |
| **original edges** | Antigenes, Nearchus, Satraps<br>**Antigenes, Nearchus, Tlepolemus** | 0.74<br>0.26 | Stefanina, Thessaloniki<br>**Nea Madytos, Vrasna** | 0.75<br>0.25 | Skra, Kilkis<br>**Greek Macedonia, Despotate of Epirus** | 0.38<br>0.01 |
| **similarity-hyperedge** | Antigenes, Nearchus, Satraps<br>**Antigenes, Nearchus, Tlepolemus** | 0.64<br>0.36 | Stefanina, Thessaloniki<br>**Nea Madytos, Vrasna** | 0.67<br>0.33 | Skra, Kilkis<br>**Greek Macedonia, Despotate of Epirus** | 0.26<br>**0.03** |
| **DHnode-in-out-degree** | Antigenes, Nearchus, Satraps<br>**Antigenes, Nearchus, Tlepolemus** | 0.69<br>0.31 | Stefanina, Thessaloniki<br>**Nea Madytos, Vrasna** | 0.78<br>0.22 | Skra, Kilkis<br>**Greek Macedonia, Despotate of Epirus** | 0.29<br>0.01 |
| **similarity-hyperedge<br>- DHnode-in-out-degree** | **Antigenes, Nearchus, Tlepolemus** | **0.6** | Stefanina, Thessaloniki<br>**Nea Madytos, Vrasna** | 0.58<br>**0.42** | Skra, Kilkis | 0.46 |

influenced by the quality of the graph's structure and the features used.

## ACKNOWLEDGEMENTS

This research was carried out as part of the project "Optimal Path Recommendation with Multi Criteria" (Project code: KMP6-0078997) under the framework of the Action "Investment Plans of Innovation" of the Operational Program "Central Macedonia 2014-2020" that is co-funded by the European Regional Development Fund and Greece.

## REFERENCES

[1] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[2] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[3] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 793–803.

[4] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[6] S. Xiao, S. Wang, Y. Dai, and W. Guo, "Graph neural networks in node classification: survey and evaluation," *Machine Vision and Applications*, vol. 33, pp. 1–19, 2022.

[7] B. Li and D. Pi, "Learning deep neural networks for node classification," *Expert Systems with Applications*, vol. 137, pp. 324–334, 2019.

[8] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," *arXiv preprint arXiv:1907.10903*, 2019.

[9] A. Kumar, S. S. Singh, K. Singh, and B. Biswas, "Link prediction techniques, applications, and performance: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 553, p. 124289, 2020.

[10] L. Cai, J. Li, J. Wang, and S. Ji, "Line graph neural networks for link prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5103–5113, 2021.

[11] "Wikispeedia navigation paths - SNAP: Stanford," [retrieved: February 8, 2024]. [Online]. Available: http://snap.stanford.edu/data/wikispeedia.html

[12] M. Sotiroudi, A.-S. Toufa, and C. Kotropoulos, "Central Macedonia Wikipedia dataset," [retrieved: February 8, 2024]. [Online]. Available: https://tinyurl.com/5n7yd94a

[13] "Code for WCM dataset creation," [retrieved: February 9, 2024]. [Online]. Available: https://github.com/MarthaSotiroudi/Wikipedia-Central-Macedonia-Dataset

[14] J.-B. Cordonnier and A. Loukas, "Extrapolating paths with graph neural networks," *arXiv preprint arXiv:1903.07518*, 2019.

[15] J. Jaehyeong *et al.*, "Edge representation learning with hypergraphs," in *Advances in Neural Information Processing Systems*, vol. 34. Virtual Conference: Curran Associates, Inc., 2021, pp. 7534–7546.

[16] A.-S. Toufa, C. Kotropoulos, and I. Tsingalis, "Dual hypergraph features for path inference in wikipedia links," in *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–7.

[17] A.-S. Toufa, I. Tsingalis, and C. Kotropoulos, "DualGRETEL+: Exploiting dual hypergraphs for path inference applied to navigation data," in *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics with International Participation (PCI)*, 2023, pp. 1–10.

[18] C. Kotropoulos, "Multimedia social search based on hypergraph learning," in *Graph-Based Social Media Analysis*, I. Pitas, Ed. CRC Press, 2016, vol. 39, pp. 215–273.

[19] "Wikipedia Central Macedonia article," [retrieved: February 8, 2024]. [Online]. Available: https://en.wikipedia.org/wiki/Central_Macedonia

[20] S. Auer *et al.*, "Dbpedia: A nucleus for a web of open data," in *International Semantic Web Conference*. Springer, 2007, pp. 722–735.

[21] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*. Citeseer, 2003, pp. 1–4.

[22] "Wikispeedia Paths & Dual Hypergraph Features repository," [retrieved: February 9, 2024]. [Online]. Available: https://github.com/asrtoufa/wikispeedia-paths-dual-hypergraph-features/tree/main

[23] M. Kempton, "Non-backtracking random walks and a weighted Ihara's theorem," *arXiv preprint arXiv:1603.05553*, 2016.

[24] "Gephi - The Open Graph Viz Platform," [retrieved: February 8, 2024]. [Online]. Available: https://gephi.org/

 ISBN: 978-1-68558-138-1 