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, GREECE
2024 InfoSys 2024 & InfoWare 2024
Int. Conf. on Advances in Databases, Knowledge, and Data Applications
costas@csd.auth.gr

March 13, 2024

Presenter's bio

Constantine Kotropoulos received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki.
He is currently a Full Professor in the Department of Informatics at the Aristotle University of Thessaloniki. He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA during the academic year 2008-2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993. He has co-authored 69 journal papers, 221 conference papers, and contributed 9 chapters to edited books in his areas of expertise.
He is co-editor of the book "Nonlinear Model-Based Image/Video Processing and Analysis" (J. Wiley and Sons, 2001). His current research interests include forensics; audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques; and human-centered multimodal computer interaction.
Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a senior member of the IEEE and a member of EURASIP, IAPR, and the Technical Chamber of Greece. He was a Senior Area Editor of the IEEE Signal Processing Letters and he has been a member of the Editorial Board of the journals: Advances in Multimedia, International Scholar Research Notices, Computer Methods in Biomechanics & Biomedical Engineering: Imaging & Visualization, Artificial Intelligence Review, MDPI Imaging, MDPI Signals, and MDPI Methods and Protocols.
Prof. Kotropoulos served as Track Chair for Signal Processing in the 6th Int. Symposium on Communications, Control, and Signal Processing, Athens, 2014; Program Co-Chair of the 4th Int. Workshop on Biometrics and Forensics, Limassol, Cyprus, 2016; Technical Program Chair of the XXV European Signal Processing Conf., Kos, Greece, 2017; Technical Program Chair of the 5th IEEE Global Conf. Signal and Information Processing, Montreal, Canada, 2017; General Chair of the 2022 IEEE 14th Image, Video and Multidimensional Signal Processing Workshop, Nafplio, Greece; Technical Program Chair of the 2023 IEEE International Conf. on Acoustics, Speech, and Signal Processing, Rhodes, Greece.

Outline

1. Introduction
2. Methodology
   - Dataset Creation
   - Feature Extraction
   - Path Extrapolation Using GRETEL
3. Experimental Evaluation
4. Concluding Remarks

- Graph structures capture relationships within data, enabling advanced analysis through Graph Neural Networks (GNNs).
- GNN applications range from node classification to link prediction, with a focus on predicting agents' trajectories over graphs.
- The effectiveness of GNNs depends on the quality and structure of the underlying graph.

- Creation of the Wikepedia Central Macedonia (WCM) Dataset, a new dataset focusing on Central Macedonia, designed to mimic human Wikipedia navigation patterns. This dataset facilitates the study of path inference in graph structures.
- Application of Dual Hypergraph Transformation (DHT) to capture complex interactions and enhancing feature extraction for more accurate path inference.
- Utilization of the GRETEL and the Dual GRETEL models for path extrapolation. These models are assessed on the WCM dataset to explore their efficacy in different graph densities.

- Construction of the dataset through a crawling process on Wikipedia, focusing on articles about Central Macedonia.
  - Initiation of the process from the article on Central Macedonia, collecting external links from each visited article.
- Random selection of next articles from the set of external links, applying validity criteria to maintain the relevance of the link.
  - Exclusion of titles containing specific terms (such as Talk, User, etc.) using the `is_valid_title` function.
- A total of 3000 paths were created, represented in a graph G with m nodes and n edges.

Methodology

Dataset Creation

- Construction of Two Graph Types:
  - Dense Graph: Uses a modified path selection process targeting the first five external links, leading to a denser graph structure with 912 nodes, 1311 edges, and 3000 paths.
  - Sparse Graph: Utilizes a broader selection from all external links, creating an extensive graph with 7307 nodes, 10612 edges, and 3000 paths.
- Paths are documented as trajectories, each with a unique identifier, using tensor manipulation for node indexing.
- Graphs are represented using the Graph Markup Language, differentiating between Dense and Sparse Graph structures for comprehensive analysis.

Methodology

Article Categorization

- Identifies content, trends, and patterns within paths.
- Utilizes dynamic online querying of Wikipedia articles via DBpedia's SPARQL endpoint to determine their semantic type.
- It is based on DBpedia's ontology, enriching the dataset for research, analytical, and educational uses.
- In cases where an explicit type is not obtained or querying errors occur, articles are classified under `subject.General`.

Methodology

Primal Feature Extraction in GRETEL

- Conducted to leverage semantic information from article content and captures complex graph interactions.
- Node feature vector includes `in/out degree`, with a length of 2.
- Edge feature vector contains `TF-IDF score` for semantic similarity and the number of times a link was clicked `nof`.

Methodology

Dual GRETEL

- Initial setup with graph G having n nodes and m edges, represented by feature matrices $\boldsymbol{F}$ and $\boldsymbol{E}$.
- Incidence matrix $\boldsymbol{M}$ differentiates directed and undirected graphs, indicating node-edge relationships.
- DHT interchanges the roles of nodes and edges
  $G = (\boldsymbol{F}, \boldsymbol{M}, \boldsymbol{E}) \rightarrow G^* = (\boldsymbol{E}, \boldsymbol{M}^\top, \boldsymbol{F})$
- The DHT is reversible, preserving structural and feature integrity.

- Two new features: `similarity-hyperedge` and `DHnode-in-out-degree`.
- `similarity-hyperedge` uses cosine similarity between incidence row vectors for semantic analysis.
- `DHnode-in-out-degree` assesses in/out degrees of dual hypergraph nodes, normalized by $D_{\max}$ for comparability.

$$\text{Normalized In/Out-Degree } \left(v_i^*\right) = \frac{\text{In/Out-Degree } \left(v_i^*\right)}{D_{\max}} \tag{1}$$

- These features are critical for capturing node relationships in both dense and sparse graph contexts.

- Focuses on predictive path analysis using the GRETEL model, estimating the conditional likelihood of path suffixes.
- Graph $G = (\mathcal{V}, \mathcal{E})$ represents the traversal landscape with $n = |\mathcal{V}|$ nodes and $m = |\mathcal{E}|$ edges.
- Agent's trajectory is represented as prefix p (traversed nodes) and suffix s (potential future nodes) for prediction horizon h.
- Agent position is encoded by sparse vector $\boldsymbol{x}_t$, normalized to unit sum, indicating likelihood of presence at node v_i.

Methodology

Path Extrapolation Using GRETEL

- GRETEL employs MLP for edge weight computation, integrating node features $\boldsymbol{f}_i$, $\boldsymbol{f}_j$ and edge features $\boldsymbol{f}_{i\to j}$:

$$z_{i\to j} = \mathrm{MLP}(\boldsymbol{c}_i, \boldsymbol{c}_j, \boldsymbol{f}_i, \boldsymbol{f}_j, \boldsymbol{f}_{i\to j}) \tag{2}$$

- $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are the pseudo-coordinates of the sender and the receiver node, respectively that are computed using a GNN of K layers.
- The original graph is modified to encode the directionality of the observed *prefix* path. A non-backtracking walk is applied to the modified graph generating candidate suffixes.
- The output of the model is the conditional likelihood $\Pr(s \mid h, p, G)$.
- The model predicts future positions $\hat{\boldsymbol{x}}_{t+h}$ using non-backtracking walks, leveraging the graph's structure and edge directionality.

Experimental Evaluation

Datesets and methods

Table 1: Dataset Characteristics

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

- Experiments are conducted on the Sparse Graph, the Dense Graph, and the Wikispeedia dataset.
- GRETEL and Dual GRETEL are utilized for path prediction, comparing original edge features against features extracted from the dual hypergraph.

Experimental Evaluation

Performance metrics

- 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.
- Precision top5 evaluates how often the correct next step appears within the model's first five predictions.

Experimental Evaluation

- Gephi's modularity class algorithm is used to delineate clusters within the network, grouping nodes based on the density of connections. The method highlights clusters where nodes are more interconnected among themselves than with the rest of the network.
- The size of each node is scaled according to its degree, allowing for the quick identification of nodes with a high number of connections, indicating their centrality and importance within their clusters.
- Visible labels on nodes are assigned based on their degree, emphasizing the most influential topics within each cluster of the Wikipedia network.

Experimental Evaluation

Sparse Wikipedia Graph Visualization

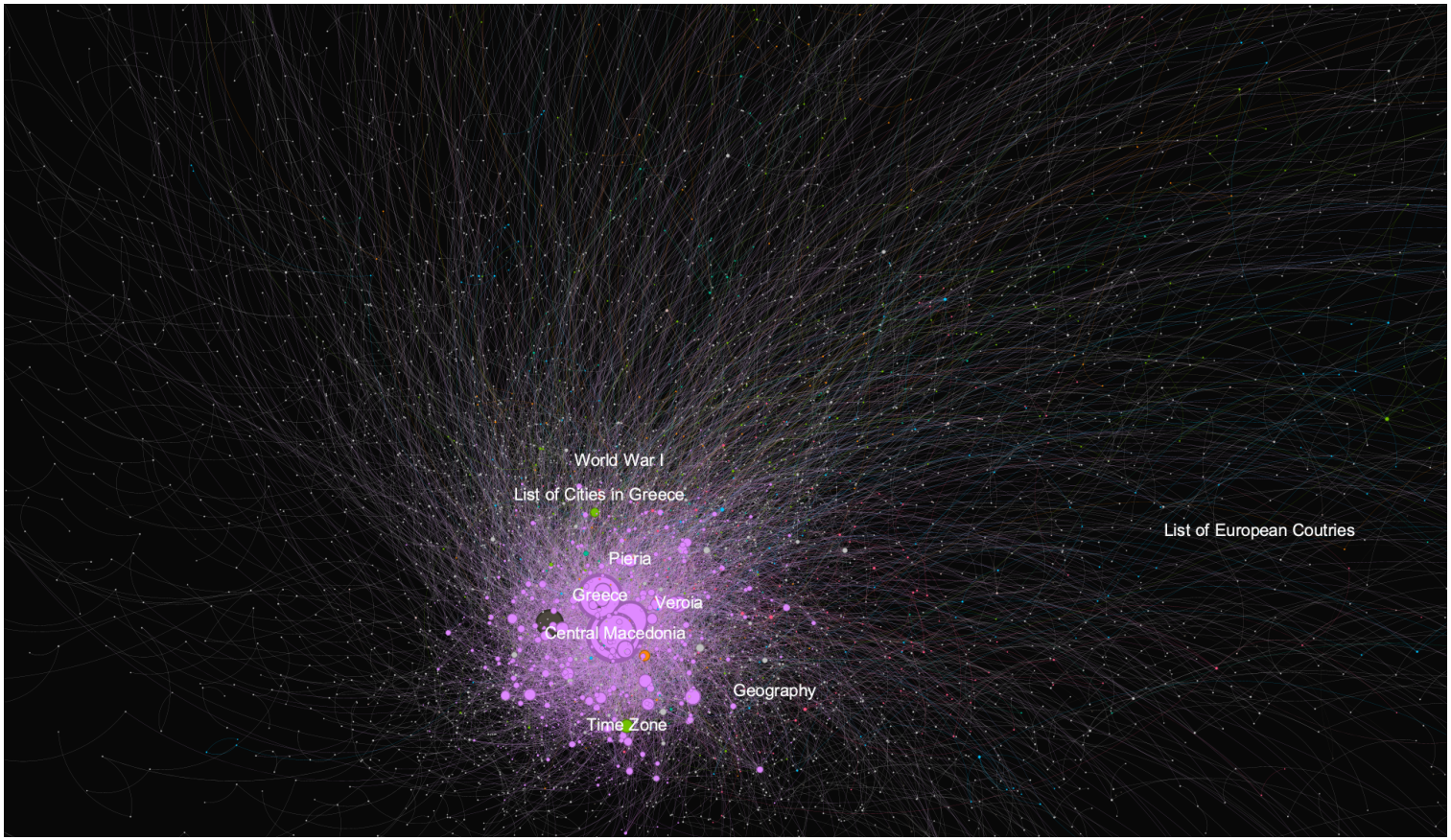

Figure 1: Sparse Wikipedia Graph. Illustrates the network formed from the Central Macedonia article, showcasing a primary cluster and smaller thematic clusters.

Experimental Evaluation

Dense Wikipedia Graph Visualization

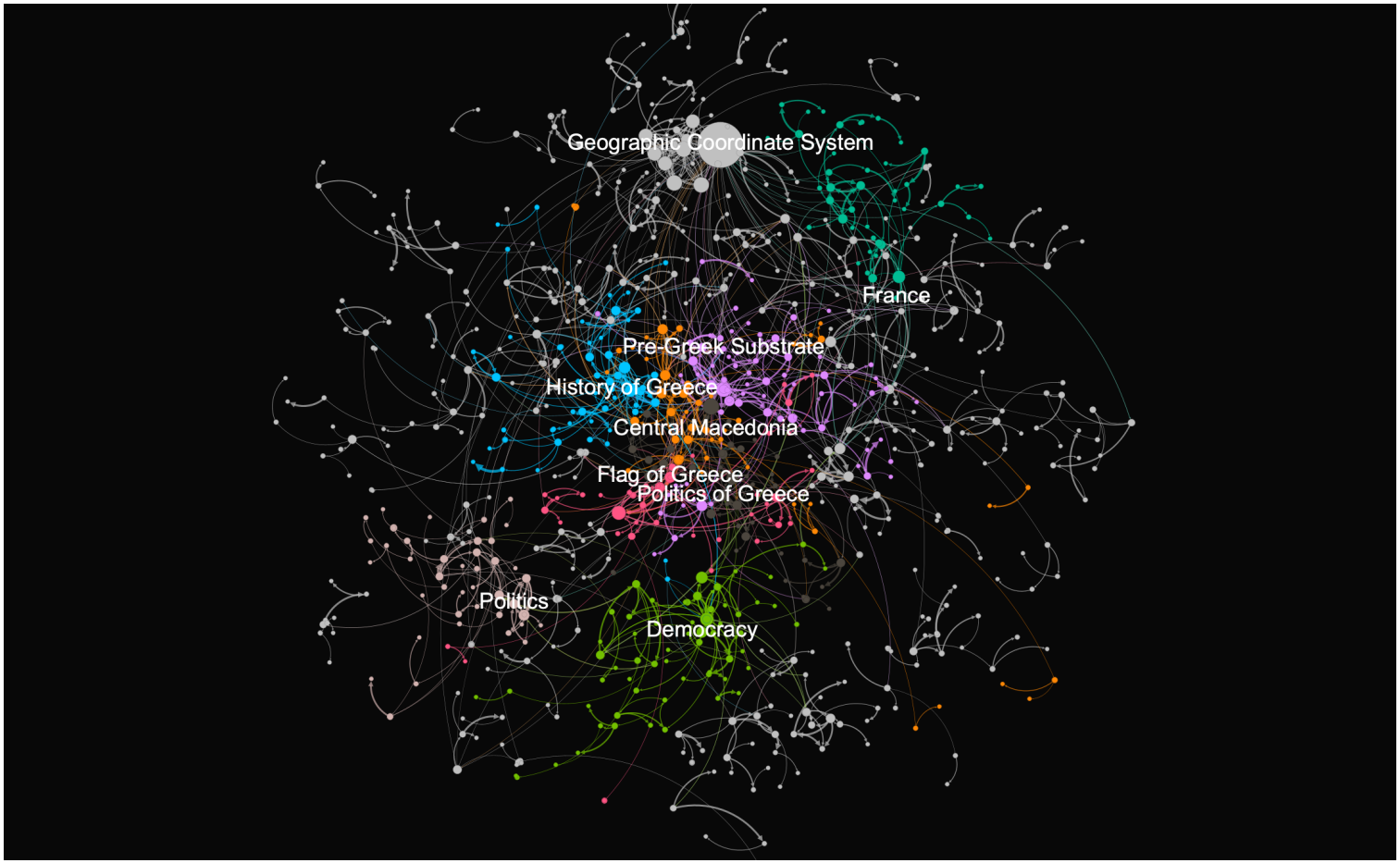

Figure 2: Dense Wikipedia Graph. The graph shows clustering around the Central Macedonia article, with nodes like 'History of Greece' indicating thematic connections.

Experimental Evaluation

Wikispeedia Graph Visualization

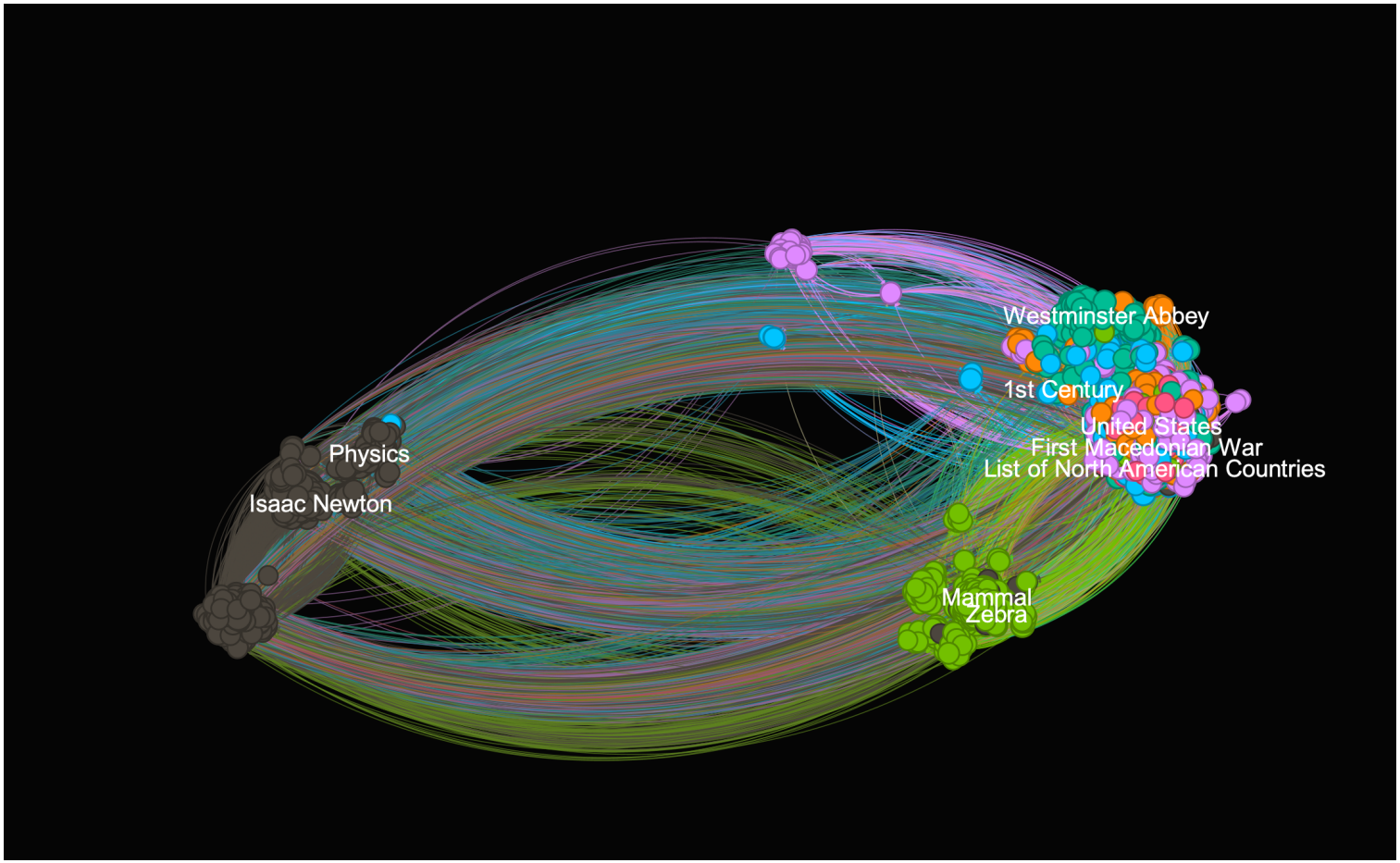

Figure 3: Wikispeedia Graph. Shows uniformly sized nodes and thematically organized clusters, such as scientific inquiry and English history.

Experimental Evaluation

Path Predictions

Table 2: 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,** **History of Macedonia, Craterus**		**Volvi, Egnatia,** **Thessaloniki, Arethousa**		**Thessaloniki, Greek National Road,** **Evzonoi, Axioupoli**	
true suffix	**Antigenes, Nearchus, Tlepolemus**		**Nea Madytos, Vrasna**		**Greek Macedonia, Despotate of Epirus**	
original edges	Antigenes, Nearchus, Satraps **Antigenes, Nearchus, Tlepolemus**	0.74 0.26	Stefanina, Thessaloniki **Nea Madytos, Vrasna**	0.75 0.25	Skra, Kilkis **Greek Macedonia, Despotate of Epirus**	0.38 0.01
similarity-hyperedge	Antigenes, Nearchus, Satraps **Antigenes, Nearchus, Tlepolemus**	0.64 0.36	Stefanina, Thessaloniki **Nea Madytos, Vrasna**	0.67 0.33	Skra, Kilkis **Greek Macedonia, Despotate of Epirus**	0.26 0.03
DHnode-in-out-degree	Antigenes, Nearchus, Satraps **Antigenes, Nearchus, Tlepolemus**	0.69 0.31	Stefanina, Thessaloniki **Nea Madytos, Vrasna**	0.78 0.22	Skra, Kilkis **Greek Macedonia, Despotate of Epirus**	0.29 0.01
similarity-hyperedge **- DHnode-in-out-degree**	**Antigenes, Nearchus, Tlepolemus**	0.6	Stefanina, Thessaloniki **Nea Madytos, Vrasna**	0.58 0.42	Skra, Kilkis	0.46

- Utilizing hypergraph features significantly increases the model's probability of accurately identifying the correct path.

Experimental Evaluation

Sparse Graph Experiments

Table 3: 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 ± 0.0044	68.76 ± 0.0019	68.99 ± 0.0064	69.71 ± 0.0038
choice accuracy	51.18 ± 0.0011	38.69 ± 0.0042	39.60 ± 0.0082	39.24 ± 0.0090
precision top1	66.65 ± 0.0050	66.71 ± 0.0012	66.65 ± 0.0025	67.14 ± 0.0045
precision top5	80.62 ± 0.0019	80.62 ± 0.0019	80.68 ± 0.0023	80.98 ± 0.0036

- Dual GRETEL predicts the correct target with a notable probability of $69.71 \pm 0.0038\%$.
- GRETEL's accuracy in choosing the next step is measured at $51.18 \pm 0.0011\%$.

Experimental Evaluation

Dense Graph Experiments

Table 4: 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 ± 0.0021	19.1007 ± 0.0004	18.8741 ± 0.0033	19.0980 ± 0.0026
choice accuracy	48.0602 ± 0.0135	27.8261 ± 0.0084	29.8662 ± 0.0096	29.5318 ± 0.0086
precision top1	0.001 ± 0.0023	19.8995 ± 0.0074	18.0904 ± 0.0075	20.5025 ± 0.0067
precision top5	0.2513 ± 0.0012	83.2161 ± 0.0088	82.8141 ± 0.0258	83.8694 ± 0.0112

- Dual GRETEL significantly enhances its performance on the WCM dense graph, achieving a precision top5 score of 83.8694±0.0112%.
- Despite the advantages in precision tasks, the dense graph's intricate connections pose challenges, such as potential overfitting and increased noise.

Experimental Evaluation

Wikispeedia Dataset Experiments

Table 5: 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 ± 0.1	6.74 ± 0.1	6.44 ± 0.2	6.2 ± 0.1
choice accuracy	22.16 ± 0.4	23.2 ± 0.1	22.88 ± 0.1	21.86 ± 0.4
precision top1	11.6 ± 0.2	12.7 ± 0.1	12.14 ± 0.1	11.66 ± 0.3
precision top5	30.1 ± 0.1	30.14 ± 0.1	30.02 ± 0.05	30 ± 0.09

- On the WIKISPEEDIA graph, Dual GRETEL also exhibits enhanced performance with a precision top5 score of $30.14 \pm 0.1\%$.
- In denser graphs, the rise in possible paths and the presence of noise from less relevant connections are major factors that make accurate path prediction more challenging.

Concluding Remarks

- The addition of Dual Hypergraph features improves GRETEL performance significantly in dense graphs.
- In sparse graphs, despite lower connectivity, the GRETEL model predicts better the correct target and exhibits better accuracy in choosing the next target, compared to dense graphs.
- The GRETEL model shows markedly better performance in the dense graph than the WIKISPEDIA graph, especially after the application of hypergraph features.
- As graph density increases, the path prediction model's performance decreases due to more paths introducing challenges in accuracy. A denser network may add more noise with weaker links, potentially misleading the path prediction algorithm.

Github Code

- The GitHub repository for the dataset creation process can be found at: `https://github.com/MarthaSotiroudi/Wikipedia-Central-Macedonia-Dataset`
- The GitHub repository for the Dual GRETEL code and experiments can be found at: `https://github.com/asrtoufa/wikispeedia-paths-dual-hypergraph-features/tree/main`

Or by scanning the QR codes:

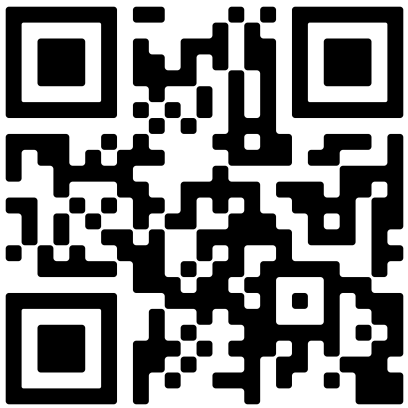

Figure 4: Dataset Creation Process

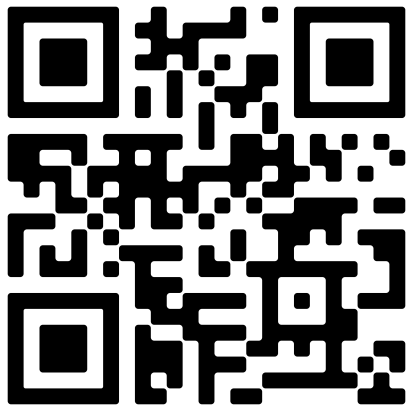

Figure 5: Dual GRETEL Code and Experiments

Acknowledgments

European Union
European Regional
Development Fund

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.

Acknowledgments

European Union
European Regional
Development Fund

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.

Thank you!

Any Questions?