Evolution and Link Prediction of the Wikipedia Network

Zecheng Zhang
zecheng@stanford.edu
Computer Science
Stanford Univeristy

Yuan Shi
yuanshi@stanford.edu
Management Science and Engineering
Stanford Univeristy

Xinwei He
xhe17@stanford.edu
Computer Science
Stanford Univeristy

ABSTRACT

As the best-known open source online encyclopedia, Wikipedia contains millions of articles connected to one another by the *wikilinks*, forming a giant network of human knowledge. In this paper, we utilize the newly released *WikiLinkGraph* dataset to analyze the evolution of the Wikipedia network in multiple languages in the form of 18 yearly snapshots. The first part of our project involves a close examination of various macroscopic properties and how they change overtime. We find that although the Wikipedia network demonstrates many similarities to other well-explored social networks, it does not follow a clear densification power law pattern like many other social networks do. Next, we tackle the link prediction task on this dataset first using common heuristics based on local static features, and proceed to create our own algorithms that incorporate temporal information and sub-graph level structural information. Specifically, we designed and trained a temporal logistic regression model and a GNN-based model using the data, and both models achieved significant out-performance compared to the baselines.

KEYWORDS

knowledge graph, temporal graph, link prediction, graph neural network

1 INTRODUCTION

Wikipedia is the best-known and largest online collaborative knowledge base. Created and edited by users around the world, its articles contain hundreds of millions of hyperlinks (*wikilinks*) connecting subjects to other Wikipedia pages, forming a giant semantic network of human knowledge. Understanding the evolution of the rich human-generated semantic content of Wikipedia will not only help Wikipedia to improve its content organization and recommendation, but will also allow us to peek into the dynamic process of human knowledge creation and storage.

In this paper, we explore the evolution of the Wikipedia network by tapping into a newly released Wikipedia dataset, *WikiLinkGraph* [3], which contains yearly snapshots of the network for nine largest language editions through 2001 to 2018. Then we extend several link prediction models to better capture the temporal and structural information in that Wikipedia dataset.

The first part of the project focuses on examining the macroscopic graph structure over the past 18 years and comparing the results with existing theories of real world dynamic networks.

The second part of the project tests a range of link prediction heuristics common for social networks on the knowledge network, and then extends these heuristics by using the rich temporal and structural information offered by the dataset to develop machine-learning based link prediction algorithms tailored to the Wikipedia network.

We arrange the paper as follows. Section 2 outlines the problems we aim to address and provides necessary network definitions. Section 3 provides a survey of related works on large-scale and evolution network analysis and link prediction. An overview of the *WikiLinkGraph* dataset is presented in Section 4. Section 5 and 6 delve into the first and second part of the project respectively, explaining our approach to the problems in detail. Experimental results for link prediction are presented and discussed in Section 7. Finally, we summarize our findings and comment on current limitations and possible directions for future research in Section 8.

2 PROBLEM STATEMENT AND DEFINITIONS

This project aims to apply and develop suitable network analysis tools to examine dynamic Wikipedia data in the form of yearly snapshots, in order to capture the evolving macroscopic picture and perform the task of link prediction. We believe a better understanding of these problems would help to inform content organization and recommendation for knowledge networks and reveal the dynamics of knowledge creation and storage.

Formally, we approach the above mentioned problem by testing the following three hypotheses:

Hypothesis 1. The Wikipedia network exhibits similar macroscopic properties as other well-explored real world networks: highly skewed degree distribution, high clustering coefficient compared to a random graph, and has one giant weakly connected component. We expect these characteristics to stay stable across different snapshots.

Hypothesis 2. The evolution of Wikipedia follows the densification power law and shrinking diameters [7]: The average node degree of the graph is increasing (and hence with the number of edges growing super-linearly in the number of nodes) and follows a power-law pattern; effective diameters of the graph tend to decrease over time.

Hypothesis 3. Local network structural features of Wikipedia have predictive power over link formation between the pages; further incorporating subgraph level structures and temporal features improves the predictive power.

We further introduce the definitions of the network and necessary network features of the network analysis in the following part of this section.

We treat the Wikipedia link network as a dynamic, directed graph $G_t = (N_t, E_t)$, where $t \in \{1, 2, \ldots, 18\}$. Each node $i_t \in N_t$ is a Wikipedia page that is active at the time t, represented by its unique page ID matching a unique page title. A directed edge $(i, j)_t \in E_t$ represents a hyperlink in page i directing to page j that is active at time t. Individual nodes and edges may be mapped to different feature vectors v_t based on its local graph structures as t varies.

We use the following standard definitions for graph features.

Definition 2.1. **In-degree** of a node i is the number of edges pointing towards node i from its neighbours, and **out-degree** of i is the number of edges pointing from i to its neighbours.

Definition 2.2. The **clustering coefficient** of node i, C_i, is defined as

$$C_i = \frac{2e_i}{k_i(k_i - 1)},$$

where e_i is the number of edges between the neighbors of node i and k_i is the total degree of node i (in- and out-degree).

Definition 2.3. The **largest weakly connected component** (WCC) is defined as the largest set where any two nodes can be joined by a undirected path.

Definition 2.4. For each natural number d, let $g(d)$ denote the fraction of connected node pairs whose undirected shortest connecting path in a graph G has length at most d. Let D be an integer for which $g(D-1) < 0.9$ and $g(D) \geq 0.9$. Then the graph G has the (integer) **effective diameter** D. [16]

3 RELATED WORK

The analysis of the Wikipedia network involves previous works on large-scale network structure and evolution. For the link prediction task we utilize and extend some existing state-of-art approaches. In what follows we give a brief overview of past works that are related to our approaches.

Large-Scale Network Structure and Evolution

An extensive range of tools have been built to study large scale, dynamic real-world graphs. [12] provided a detailed study of comparison between real-world network structures with those of random graphs, and pointed out some measurable discrepancies that indicated the presence of additional social structure not captured by the random graph. According to [7], on the other hand, cast their attention on the evolution dynamics of these social networks and identified shrinking diameters and densification as their key dynamic characteristics. The forest fire model was proposed to describe network evolution with these characteristics.

Despite the abundance of past works in the field of large-scale network analysis, most of these papers have focused on networks with a strong social component instead of knowledge networks. Whether the same characteristics and dynamics seen in social networks are applicable to knowledge networks, however, is still not obvious.

Among the limited set of literature that studied knowledge networks are [1], which examined a 2005 Wikipedia snapshot and detected cultural bias based on global page ranking given by HITS and PageRank score, and [9] which studied the community structure of the a 2009 Wikipedia snapshot using the Girvan-Newman algorithm [2], a method based on greedy maximization of modularity.

Existing works of this kind have mostly focused on single, static snapshots of early data and studied only a small subset of the current Wikipedia network, and they often faced the challenge of messy large-scale data with redundant, auto-generated links and cliques that are hard to process and interpret. It is questionable whether these results remain valid today.

Link Prediction

The problem of link prediction in time-evolving networks has received significant attention, both in the subject of link addition [10] and link removal [4]. Liben-Nowell and Kleinberg [8] are among the first to propose and test a range of methods for link prediction in social networks, using heuristics including the number of common neighbors, Jaccard similarity, preferential attachment etc. More recently, graph neural network based approaches, such as [18], are proposed to unify these heuristics and enrich them with mesoscopic structural information through generation of subgraph embeddings.

In addition to heuristics and algorithms based structural features, [15] designed temporal node features using a series of graph snapshots and applied them in the task of temporal link prediction.

Although many of these techniques have proved to be useful in many real-world networks, application on knowledge networks such as Wikipedia remains rare. Preusse [14] investigated evolution of links in non-English Wikipedia graphs using basic heuristics including node degree, joint degree, age of edge and nodes, etc. The performance of these heuristics was evaluated using area under the ROC Curve (AUC), standard for link prediction tasks [6] with acceptable performance.

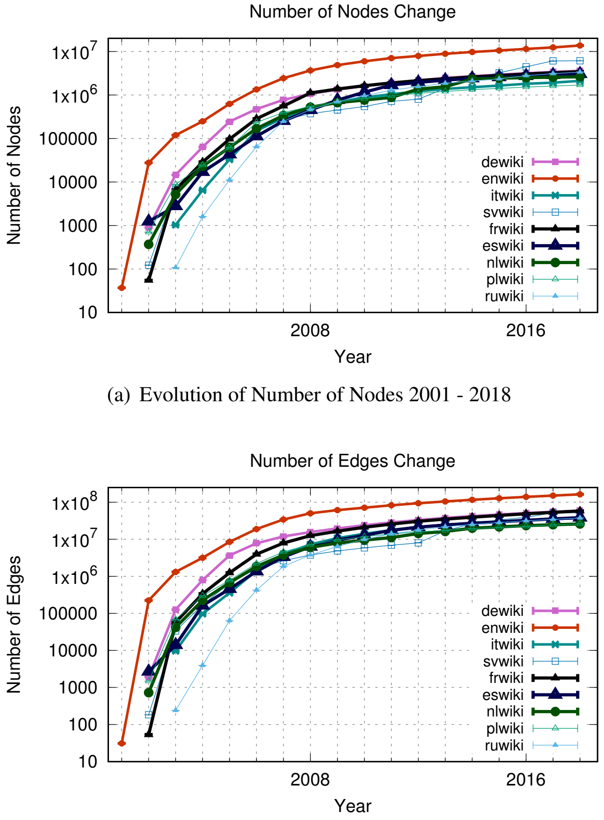
Beyond these simple heuristics based on one-dimensional features, more sophisticated, multi-dimensional temporal features have not been explored on Wikipedia data to our knowledge.

In general, few people have made a comprehensive analysis on a relatively complete Wikipedia link network dataset with temporal information. The problem of link prediction in large-scale Wikipedia networks is also seldom studied with rigor, due to limited data availability and computing power constraints. Our analysis will take a first step into this blank space by tapping into the newly released *WikiLinkGraph* dataset and extend the link prediction task to that dataset.

4 THE WIKIPEDIA NETWORK DATASET

The recently published dataset *WikiLinkGraphs* [3] is one of the most complete Wikipedia network with snapshots information. It is a complete dataset of the network of internal Wikipedia links for the 9 largest language editions. The dataset contains yearly snapshots of the networks and spans 18 years, from the creation of Wikipedia in 2001 to March 1st, 2018.

In contrast to previous Wikipedia data obtained through database dumps containing links automatically generated by templates, *WikiLinkGraphs* parsed the data so that only links explicitly added by editors in the text of the articles are included. The result is a cleaner dataset where the links provide a more trustful representation of semantic relations between concepts, and potential cliques and anomalous patterns generated by transcluded links are avoided.



(a) Evolution of Number of Nodes 2001 - 2018

(b) Evolution of Number of Edges 2001 - 2018

Figure 2: Evolution of the number of nodes and edges for Wikipedia datasets in 9 different languages over 18 years.

date	de		en	
	N	E	N	E
2001-03-01	0	0	37	31
2002-03-01	900	1,913	27,654	223,705
2003-03-01	14,545	126,711	118,946	1,318,655
2004-03-01	63,739	794,561	248,193	3,170,614
2005-03-01	244,110	3,659,389	624,287	8,505,195
2006-03-01	474,553	7,785,292	1,342,642	18,847,709
2007-03-01	775,104	11,946,193	2,425,283	34,219,970
2008-03-01	1,063,222	15,598,850	3,676,126	50,270,571
2009-03-01	1,335,157	19,607,930	4,848,297	61,318,980
2010-03-01	1,603,256	23,834,140	5,937,618	71,024,045
2011-03-01	1,879,381	28,457,497	7,027,853	82,944,163
2012-03-01	2,163,719	33,036,436	7,922,426	93,924,479
2013-03-01	2,461,158	37,861,651	8,837,308	105,052,706
2014-03-01	2,712,984	42,153,240	9,719,211	116,317,952
2015-03-01	2,933,459	46,574,886	10,568,011	127,653,091
2016-03-01	3,155,927	50,904,750	11,453,255	139,194,105
2017-03-01	3,372,406	55,184,610	12,420,400	150,743,638
2018-03-01	3,588,883	59,535,864	13,685,337	163,380,007

Figure 1: The exact numbers of nodes and edges for the two largest editions: English (en) and German (de).

Figure 2 gives an overview of the evolving number of nodes and edges across snapshots for multiple languages on a log scale, and Figure 1 presents the exact statistics for the two largest editions: English and German. In practice, we further parsed the data into graph objects for individual snapshots with nodes represented by integer indices, and separately stored dictionaries that match the indices to respective text page titles.

Hence, in this paper we utilize the *WikiLinkGraphs* as our analysis and examination dataset.

5 THE EVOLVING MACROSCOPIC PICTURE

We first analyze the Wikipedia network in an evolving macroscopic picture. We provide our approaches and analysis in this section.

Approaches

In order to investigate our first hypothesis that the Wikipedia network exhibits similar macroscopic properties as other well-explored real-world networks, we compute the following statistics for each snapshots:

- The out-degree distribution of the entire snapshot;
- The average clustering coefficient from 10,000 randomly sampled nodes for each snapshots;
- The size of the largest weakly connected component.

Secondly, to test our second hypothesis that the evolution of Wikipedia follows the densification power law and shrinking diameters, we compute the following statistics for each snapshot:

- The effective diameter;
- The average node degree (in-degree + out-degree);
- Log-log regression of the number of edges versus the number of nodes.

These statistics are analyzed across different snapshots and multiple languages, and the results are compared with the common network properties and evolution dynamics as stated in Hypotheses 1 and 2.

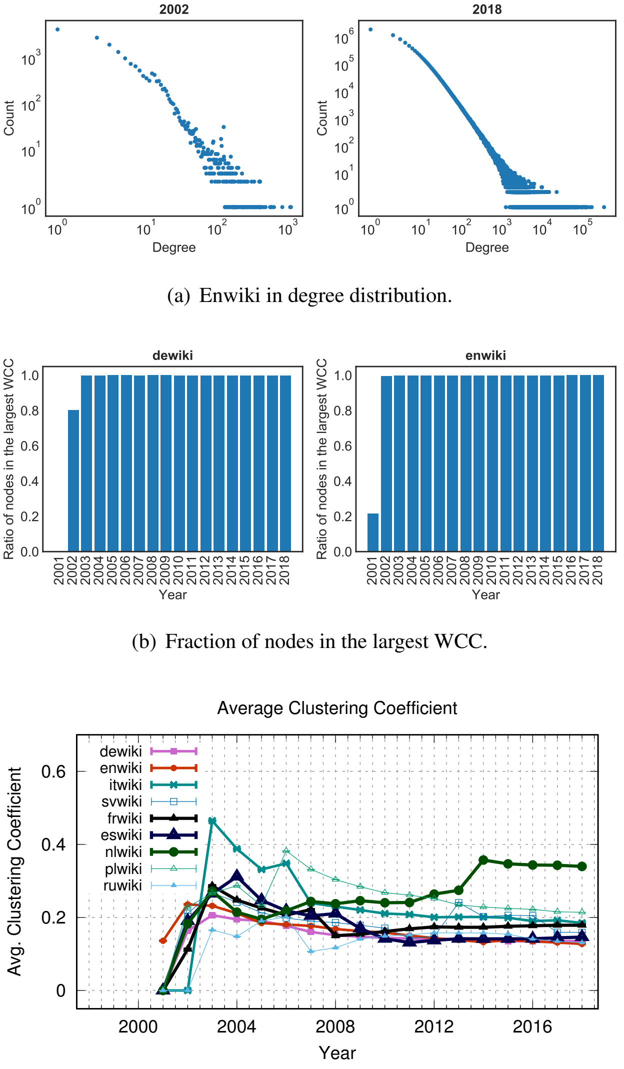
Results

Macroscopic Properties of Wikipedia.

The degree distributions, the size of the largest weakly connected component and average coefficients across different snapshots are presented in Figure 3.

Network	Degree Distn	Ave. CC	Giant WCC
MEDLINE	highly skewed	0.07	93%
MSN	highly skewed	0.11	99.9%
PPI Network	highly skewed	0.12	81%
Wikilink(en)	**highly skewed**	**0.02**	**99.9%**

Table 1: Macroscopic properties of Wikipedia vs. other real-world networks



(a) Enwiki in degree distribution.

(b) Fraction of nodes in the largest WCC.

(c) Average clustering coefficient across 18 years.

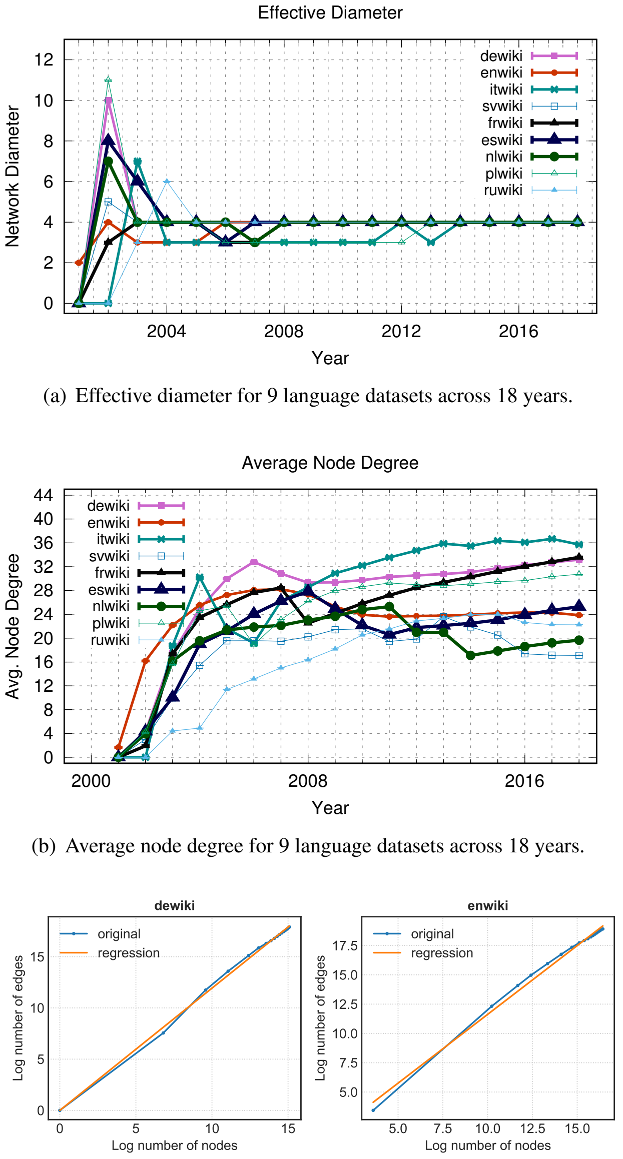
Figure 3: The Wikilink network has a highly skewed degree distribution, a giant connected component and a high clustering coefficient since the early stage of its formation.

Table 1. compares these statistics with 3 other real-world networks: MEDLINE [13], PPI Network and the MSN messaging network discussed in class. The results corroborate with Hypothesis 1 that the macroscopic properties of Wikipedia are similar to other real world networks: highly skewed, heavy-tailed degree distribution, high clustering coefficient and has one giant weakly connected component. The clustering coefficient is relatively lowed compare the the others, but still much higher than that of a random graph. Moreover, these characteristics emerged at a very early stage of Wikipedia (around 2002) and stayed consistent afterwards.

This result demonstrates that the interconnected nature of human knowledge resembles what we usually see in real-world social and biological networks. This could imply there are potential similarities in the underlying graph generation process, such as preferential attachment or community guided attachment [7].

The Dynamics of Evolution.

The dynamics of evolution characterized by the average node degree and effective diameter, on the other hand, shows interesting patterns that differ from our expectation.



(a) Effective diameter for 9 language datasets across 18 years.

(b) Average node degree for 9 language datasets across 18 years.

(c) Linear fit of log-log edges and nodes.

Figure 4: The Wikilink network demonstrates no clear densification or diameter shrinkage after the earliest formation phase. This is reflected in constant effective diameters, mixed behaviors in average node degrees and a lack of linear fit in the log-log regression of edges vs. nodes (obvious if we exclude the first data point).

As presented in Figure 4, we see mixed behaviors in the average node degrees and largely unchanged effective diameters in the past 10 years. The largest edition, enwiki, notably exhibits decreasing to constant average node degrees after 2007. While the log-log regression of number of edges and number of nodes seem to show a decent linear fit in the bottom two charts of Figure 4, it is also clear that the original plot has a concave shape with a decreasing gradient if the first data point from the earliest snapshot is excluded.

All of the three statistics suggest that ***there is no observable densification or diameter shrinkage*** in the largest Wikepedia edition after 2007.

One possible explanation for this unexpected phenomenon is that certain link formation mechanisms common in social networks may be absent in a knowledge network. A typical example would be that triadic closure through the dynamic process of ‘a friend of friend becomes friend’. As different concepts do not actively interact with one another forming new connections like people do, triadic closures in knowledge networks usually come from shifts in our perceptions of these concepts, which tends to be a slower and less common process as compared to social interaction. It is also less obvious that shifts in perceptions would necessarily make the graph denser as it does in the case of friendship formation. Additionally, missing links due to infrequent edition of existing pages could also be a possible factor.

6 LINK PREDICTION

The second part of the project addresses the challenge of link prediction. Our goal is to design a mapping from an ordered node pair, $(i, j)_t$, to a vector $v(i, j)_t \in \mathbb{R}^p$ containing in total p node level and pair level features and train a machine learning model that predicts link addition based on the mapping. Formally, out task can be defined as

$$f\ (v(i,j)_t) \rightarrow IsEdge(i,j)_{t+1} \tag{1}$$

where $IsEdge(i, j)_t$ equals 1 if an active edge (i, j) is present at time t and 0 otherwise. The temporal embedding learning problem is then an optimization problem minimizing the loss function L, e.g. the cross-entropy loss, of the task:

$$\underset{f,v}{\operatorname{argmin}}\,(L(f,v)) \tag{2}$$

Link Prediction Heuristics

We start from three common heuristics for link prediction based on static local structural features of a node pair (x, y).

(1) **The Jaccard Coefficient** (JC):

$$|\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$$

(2) **The Number of Common Neighbors** (CN):

$$|\Gamma(x) \cap \Gamma(y)|$$

(3) **Preferential Attachment** (PA):

$$|\Gamma(x|) \cdot |\Gamma(y)|$$

where $\Gamma(x)$ represents the set of neighbors of node x.

Each of the heuristics assigns a score to the node pair. The performance of these heuristics can be measured through pairwise comparison of positive and negative samples, where the positive sample should receive a higher score.

Incorporating Temporal Information

Part of our hypothesis is that temporal features may improve predictive power. To test the validity of this hypothesis, we train and compare the performance of two machine learning models: logistic regression trained on static node pair features versus logistic regression trained on node pair features from a series of consecutive historical snapshots.

For the former simple logistic regression model, node pair feature v_t is a 3-dimensional vector constructed by concatenating the 3 heuristic scores at time t only: Jaccard, Common Neighbours and Preferential Attachment. I.e.,

$$v_t = [JC(t), CN(t), PA(t)].$$

In the latter temporal logistic regression model, a new node pair feature $\tilde{v}_t$is constructed by concatenating 3 heuristic scores of the past three years altogether to form a 9-dimensional vector

$$\tilde{v}_t = \text{Concatenate}(v_t, v_{t-1}, v_{t-2}),$$

and train the logistic regression classifier on the new feature vector.

We could also have used a more general mechanism to combine the temporal information,

$$\tilde{v}_t = g(v_t, v_{t-1}, v_{t-2}),$$

where g is a general function potentially with trained parameters. Such ideas can be the subject of further investigation.

Incorporating Mesoscopic Structural Information

Recent works in graph neural networks have offered a rigorous framework that incorporates subgraph-level structural information into machine learning models for inference on networks.

We are inspired by the SEAL model introduced in [18] which transforms the link prediction task into a graph classification task that is then tackled using a graph neural network, DGCNN [17]. In the following we first explain how to transform the link prediction task to a graph classification task, then extend the SEAL framework by using GraphSage to replace DGCNN as the GNN model in the original design.

Link Prediction as Subgraph Classification. The link prediction problem defined in Equation 1 is equivalent to a subgraph classification task:

$$f\left(G'(i,j)^d_t\right) \rightarrow IsEdge(i,j)_{t+1}, \tag{3}$$

where $G'(i,j)^d_t$ is the subgraph of node i and j at time t containing nodes connected to those two nodes within distance d, and $f(\cdot)$ is the subgraph classification model. If i and j are predicted to be connected at time $t+1$, the indicator function on the right will output the label of the graph as 1 and vice versa.

Subgraph Feature Extraction via Structural Labeling. One of the most common structures found in Wikipedia is the hierarchical structure of its category trees, which can be captured by mesoscopic features. We will use the state-of-art mesoscopic feature extraction technique mentioned in [18]: structural labeling.

Define the distance function between node i and nodes x, y as

$$d(i,x) = c_1; d(i,y) = c_2 \tag{4}$$

where i is any node in the subgraph $G'(x,y)^d_t$ and $d(i,x)$ returns the distance between node i and node x, c_1 and c_2 are the distances between node i and nodes x and y tracing both incoming and outgoing edges. Hence, the structural labels for each node are in the form of a tuple.

Then, we define the mapping function m as

$$m(c_1, c_2) \longrightarrow l \tag{5}$$

where l is the structural labeling.

The mapping of labels is similar to the orbits in graphlet. For example, we first assign label 1 to nodes x and y. Then, for any node i has the mapping $m(1,1)$, we assign label 2 to the node, and label 3 to nodes with mappings $m(2,1)$ and $m(1,2)$ etc. This mapping function, proved in [18], has a perfect hashing function as

$$l = 1 + \min(d(i,x), d(i,y)) + (d/2)[(d/2) + (d\%2) - 1] \tag{6}$$

where the structural labeling l can optimally differentiating different subgraph based structural roles for each nodes.

Additional Structural Information. Apart from the subgraph based structural labeling, we also incorporate other useful structural features including in-degree, out-degree and common neighbors. We concatenate these structural features with the subgraph labels to generate our initial structural feature vectors used for GNN training.

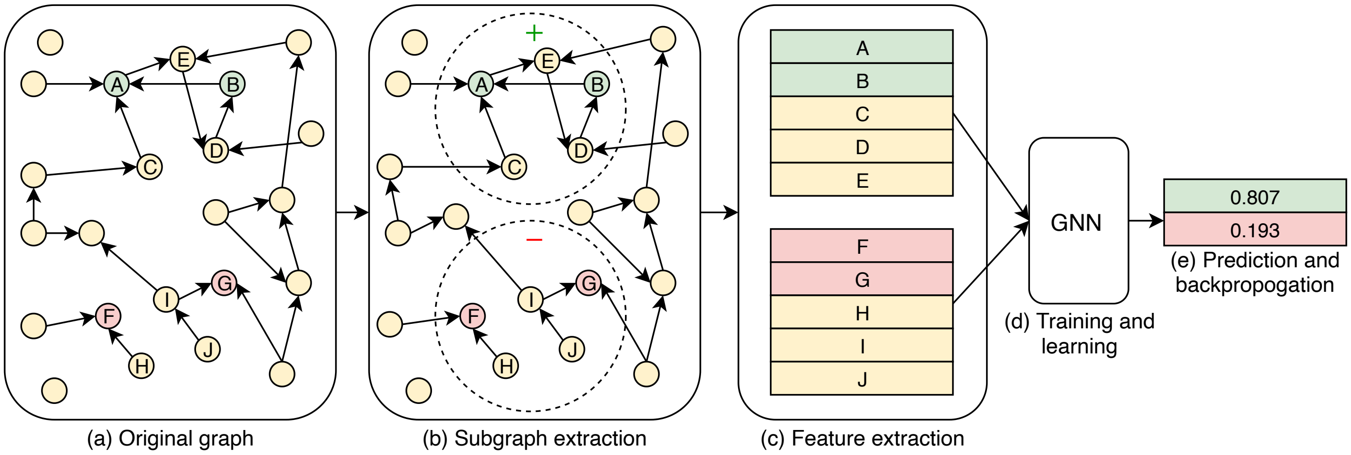


Figure 5: An overview of the GNN approach.

GraphSage as the GNN Model. To perform the graph classification task, we use GraphSAGE [5] instead of the DGCNN in the original SEAL approach. The main reason is that GraphSAGE has the state-of-art inductive learning capability, which is very helpful in the graph classification task. In this way, we extended the SEAL framework into a graph classification model for directed graphs using rich mesoscopic structural information. The general overview of the framework is shown in Figure 5.

7 EXPERIMENTS

This section presents a range of link prediction experiments we conducted using the methods proposed in the previous section, including link prediction heuristics, temporal logistic regression, and the GNN-based model. The code is available at https://github.com/RagnaroWA/wikipedia_evolution.

Experimental Settings

Dataset. We divide the data into training set and test set by a chosen pivot year, $t^* = 2014$. Snapshots dated before 2014 are used for training and the rest are used for model cross-validation.

During training, we use a snapshot with time stamp $t < t^*$ and define all node pairs (i, j) where $IsEdge(i, j)_t = 1$ as positive samples. We also randomly sample the same number of node pairs with $IsEdge(i, j)_t = 0$ as negative samples. We then test our model on a test snapshot with time stamps t' where $t' \geq t^*$. All node pairs satisfying $IsEdge(i, j)_{t'-1} = 0$ and $IsEdge(i, j)_{t'} = 1$ in the graph are considered as positive samples for the testing set, and the same number of negative samples that satisfy $IsEdge(i, j)_{t'-1} = IsEdge(i, j)_{t'} = 0$ are sampled at random.

During actual implementation, 10^4 positive samples and 10^4 negative samples are sampled at random from training and test snapshots respectively in order to overcome limitations in computational capacity.

Evaluation Metrics. We use AUC of the ROC curve as the evaluation metric for performance. The AUC value in the link-prediction setting can be interpreted intuitively as the probability that during pairwise comparison, a positive sample is given a higher score by the model than the corresponding negative sample.

Suppose out of a total number of N independent pairwise comparisons conducted, the positive sample is assigned a score higher than or equal to the negative sample for N' times, then we have

$$AUC = N'/N.$$

GNN Implementation Details. For the GNN-based method, we select GraphSAGE as the GNN model. We set the hidden layer size to 12 because the structural features usually have a dimension around 14 to 15. And we utilize the Adam optimization to update the weight matrices in backpropagation.

Results and Discussions

Performances of the three link prediction heuristics on three of the most representative editions are recorded in Table 2. The Jaccard score shows the best performance on all three languages, followed by preferential attachment.

Table 2: AUC of link prediction heuristics

Method	enwiki	dewiki	itwiki
Jaccard Coeff. (JC)	**0.614**	**0.621**	**0.632**
Preferential Attachment (PA)	0.605	0.611	0.631
Common Neighbor (CN)	0.573	0.589	0.593

The consistency in rankings provides useful information for identifying the main drivers of the link formation process in Wikipedia.

Firstly, the most predictive feature of connection between concepts is whether they share a large fraction of common neighbors. The interesting part is that the absolute number of common neighbors is less predictive than the fraction relative to the total number of neighbors. One possible explanation to this difference in performance is that a sizable proportion of links between Wikipedia pages are actually ‘weak’ links anchored in various subsections less relevant to the core concepts. For pages with a lot of out-going or incoming weak links, the absolute number of common neighbors does not always imply a strong connection. It proves to be helpful to weight the numbers by the size of the union of neighbors to more accurately reflect the strength of connection.

Secondly, the result shows that the ‘rich-get-richer’ process of preferential attachment is also common in the knowledge network. This can be explained by the hierarchical structure of the Wikipedia categorical tree[9]: high-degree nodes are likely to be higher up in the hierarchy representing more general and overarching concepts and are therefore more likely to establish connection with a wide range of other concepts.

Table 3: AUC of machine-learning based techniques

Method	AUC (itwiki)
Logistic Regression (LR)	0.730
Temporal Logistic Regression (TLR)	0.813
GNN	**0.995**

Table 3 presents the performance of machine-learning based techniques using logistic regression and GNN. We observe significant out-performance of designs with enriched temporal and structural information compared to simple logistic regression using static structural features. This corroborates with our hypothesis that temporal and sub-graph level structural features contain valuable information about the dynamics of link formation in Wikipedia. The interdependence among various concepts extends beyond the features of their immediate neighbors and is also correlated with historical trends and patterns. We believe temporal information have helped improve our prediction mainly in two ways. On the one hand, changes in features in the past years indicates that the page is undergoing active edition and probably reflects a quickly changing field or a new concept. One the other hand, old and well-established central nodes are also likely to be the subject of in-link addition through the process of preferential attachment.

Figure 6 plots the performance of all of the six techniques implemented on the Italian edition and ranks them accordingly.

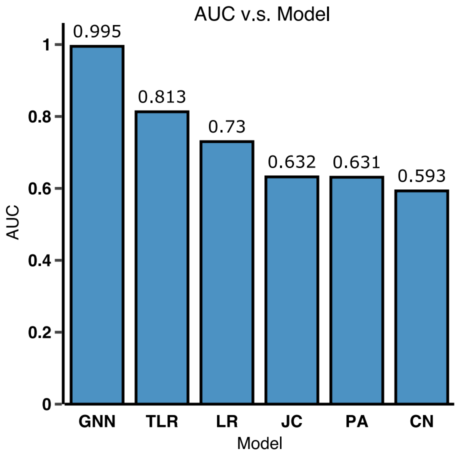


Figure 6: Plot of AUC for the 6 techniques for itwiki

8 CONCLUSIONS AND FURTHER WORK

In this paper we undertook a detailed examination of the evolution of macroscopic features of Wikipedia over the past 18 years, and designed two machine learning-based link prediction algorithms that separately incorporate temporal and subgraph-level structural information to improve task performance.

There are two key results. Firstly, we showed that while the Wikilink network resembles other real-world networks in its degree distribution and clustering phenomenon, the dynamics of its evolution as a knowledge graph is meaningfully different from most well-explored social networks as it does not demonstrate a clear trend of densification. Secondly, both of our proposed link prediction techniques, temporal logistic regression and GNN, significantly out-perform common heuristics, which reveals the amount of valuable information hidden in subgraph structures and historical patterns.

One limitation of our approach is that it trains the link prediction model on subgraphs of the giant network through sampling due to lack of computational capacity. It would be better if we could find another scalable implementation that also captures full information from the entire graph. In addition, there are many possible alternatives to incorporate temporal information beyond simple concatenation, for example, through a recurrent neural network. These problems could be the next steps of possible extentions to this project.

We expect our work to pave way for many exciting research problems in the future. For example, link prediction could help Wikipedia to improve its content organization and recommendation. Understanding the dynamics of evolution could also help us better understand how concepts grow and decay in the context of social or semantic science such as controversy mapping [11] problems.

REFERENCES

[1] F. Bellomi and R. Bonato. 2005. Network analysis for Wikipedia. In *Proceedings of Wikimania 2005, The First International Wikimedia Conference*.

[2] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[3] Cristian Consonni, David Laniado, and Alberto Montresor. 2019. WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks.

[4] Marcius Armada de Oliveira, Kate Cerqueira Revoredo, and José Eduardo Ochoa Luna. 2014. Semantic unlink prediction in evolving social networks through probabilistic description logic. In *2014 Brazilian Conference on Intelligent Systems*. IEEE, 372–377.

[5] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[6] Jure Leskovec, Jure, Faloutsos, and Christos. 2006. Sampling from large graphs. https://doi.org/10.1145/1150402.1150479

[7] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 2.

[8] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58, 7 (2007), 1019–1031.

[9] Dmitry Lizorkin, Olena Medelyan, and Maria Grineva. 2009. Analysis of Community Structure in Wikipedia. In *18th International World Wide Web Conference*. 1221–1221. http://www2009.eprints.org/191/

[10] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* 390, 6 (2011), 1150–1170.

[11] Nils Markusson, Tommaso Venturini, David Laniado, and Andreas Kaltenbrunner. 2016. Contrasting medium and genre on Wikipedia to open up the dominating definition and classification of geoengineering. *Big Data & Society* 3, 2 (2016), 2053951716666102. https://doi.org/10.1177/2053951716666102 arXiv:https://doi.org/10.1177/2053951716666102

[12] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical review E* 64, 2 (2001), 026118.

[13] M. E. J. Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98, 2 (2001), 404–409. https://doi.org/10.1073/pnas.98.2.404 arXiv:https://www.pnas.org/content/98/2/404.full.pdf

[14] Julia Preusse, Jérôme Kunegis, Matthias Thimm, Steffen Staab, and Thomas Gottron. 2013. Structural dynamics of knowledge networks. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[15] Uriel Singer, Ido Guy, and Kira Radinsky. 2019. Node Embedding over Temporal Graphs. *arXiv preprint arXiv:1903.08889* (2019).

[16] Sudhir Tauro, Christopher Palmer, Georgos Siganos, and Michalis Faloutsos. 2001. A simple conceptual model for the Internet topology, Vol. 3. 1667 – 1671 vol.3. https://doi.org/10.1109/GLOCOM.2001.965863

[17] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics (TOG)* (2019).

[18] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*. 5165–5175.