

Software Engineering Department

ORT Braude College

Capstone Project Phase A – 61998

**Citation networks evolution using Dynamic Network Embeddings**

**25-1-R-16**

**Contents**

[1 Abstract 3](#_Toc186297462)

[2 Introduction 4](#_Toc186297463)

[3 Background 6](#_Toc186297464)

[3.1 Dynamic Network Background 6](#_Toc186297465)

[3.2 Word2Vec and its inspiration for Dynamic Network Embeddings 6](#_Toc186297466)

[3.3 Citation Dynamic Network Specificities 7](#_Toc186297467)

[3.4 Random Walk 8](#_Toc186297468)

[3.5 Stochastic Gradient Descent (SGD) 9](#_Toc186297469)

[3.6 Dynamic Bernoulli Embeddings 10](#_Toc186297470)

[3.7 Dynamic Network Construction 12](#_Toc186297471)

[4 Approach 15](#_Toc186297472)

[4.1 Tracking Evolution Articles in Citation Networks 15](#_Toc186297473)

[5 Testing Plan 16](#_Toc186297474)

[6 References 17](#_Toc186297475)

# Abstract

Citation networks are modeled as a series of graphs. In these graphs, nodes represent articles, and edges indicate citations from one article to another. Each graph corresponds to a snapshot at a specific timestamp. These networks evolve with the addition of new research and references. Identifying public opinion leaders, whether articles or subjects, is essential for understanding research influence and emerging trends. Advanced programming techniques are necessary to analyze these dynamic networks effectively.

Dynamic network embeddings offer a powerful approach for analyzing evolving networks by transforming graphs into vectors within a shared vector space. By employing methods such as dynamic Bernoulli embeddings and sequential random walks, this research aims to model temporal changes effectively. Through tasks such as link prediction and evolving node detection, the approach demonstrates its capability to identify influential works and emerging research areas in citation networks.

In conclusion, dynamic network embedding methods provide a robust framework for uncovering key insights in citation networks, driving a deeper understanding of research trends and influence over time.

# Introduction

Highly cited articles are instrumental in shaping the direction of research discourse. Spanning a vast array of diverse subjects, are often published in response to the popularity of certain subjects. By detecting such works, citation networks provide valuable insights into the evolution of research priorities and the propagation of knowledge.

 As shown in Figure 1, highly cited articles serve as central nodes within citation networks, connecting diverse knowledge domains and facilitating interdisciplinary collaboration. These central articles play a crucial role in driving innovation, highlighting significant research areas, and influencing funding priorities toward emerging fields. Their impact extends beyond individual research areas, shaping the evolution of disciplines and driving transformative shifts across the broader academic landscape. Over time, their influence grows as they continue to guide future research directions and inspire new discoveries.

**Figure 1: Illustration of highly cited articles as central nodes in a citation graph, highlighting their role in connecting diverse domains of knowledge and fostering interdisciplinary collaboration.**

Static analysis methods are techniques that analyze networks by assuming a fixed structure, where nodes and edges remain unchanged over time. Traditional static analysis methods, such as DeepWalk and Node2vec models, are limited in capturing the dynamic nature of evolving networks. By treating the network as a static graph, these methods provide only a snapshot at a specific timestamp, ignoring crucial temporal variations. As a result, these approaches fail to account for critical changes such as the emergence of new influential nodes or the shifting importance of edges over time. This static perspective makes it difficult to analyze trends, predict future interactions, or understand temporal patterns in the network's structure. For instance, while DeepWalk model uses random walks to learn embeddings, it assumes a fixed structure, ignoring how nodes and edges evolve. Similarly, Node2vec model introduces flexibility in embedding generation but still lacks the ability to incorporate temporal dynamics, leading to incomplete representations of the network's evolution.

Dynamic networks address the limitations of static analysis by modeling nodes and edges that change over time. Unlike static networks, dynamic networks capture how structures evolve, providing critical insights into temporal interactions. By tracking these changes, dynamic networks enable the analysis of trends, the identification of emerging patterns, and the understanding of how relationships develop over time.

Dynamic network embedding enhances this framework by generating low-dimensional representations of nodes that capture both structural relationships and temporal changes. Using specialized techniques, these embeddings ensure proximity preservation by keeping similar nodes close in the embedding space, while also ensuring temporal continuity by minimizing unnecessary drift for stable nodes over time. This approach enables researchers to identify influential articles in citation networks, shaping funding decisions, academic hiring, and research prioritization.

Analysing citation dynamic networks over time provides significant contributions to understanding the evolution of scientific knowledge and research priorities. This allows for the detection of the identification of influential works at different stages of their lifecycle. For instance, articles that may initially receive limited attention can later emerge as central nodes, influencing research trajectories and fostering innovation. By capturing these temporal patterns, dynamic network analysis enables a more nuanced understanding of how scientific ideas propagate, evolve, and gain prominence over time.

The benefits of dynamic citation network analysis extend beyond academic insights. It provides actionable intelligence for decision-makers such as funding agencies, institutions, and policymakers. By identifying rising areas of influence and predicting future research directions, these analyses can guide the allocation of resources toward high-potential fields.

Additionally, tracking changes in the importance of nodes and edges helps in evaluating the long-term impact of research contributions, improving strategies for academic hiring, grant distribution, and collaborative initiatives. Ultimately, dynamic citation network analysis fosters a deeper understanding of scientific progress, facilitating data-driven decisions that enhance innovation and accelerate the advancement of knowledge.

# Background

## Dynamic Network Background

Dynamic networks are structured systems where the connections (edges) between entities (nodes) evolve over time, reflecting changes in relationships or interactions. Unlike static networks with fixed topologies, dynamic networks are represented by a series of time-dependent snapshots or continuous-time models, capturing the temporal variation in their structure. These changes may include the addition or removal of nodes, the strengthening or weakening of connections, or shifts in the network's overall topology. However, in citation networks, edges are only added over time and not removed, as citations represent permanent references once established.

**Definition 1 (Dynamic Networks): A dynamic network is a series of graphs Γ = {G1,...,GT } and Gt = (Vt,Et), where T is the number of graphs, Vt is a node set and Et includes all temporal edges within the timespan [St,St+1]. Each ei = (u,v,si) ∈ Et is a temporal edge between the node u ∈ Vt and the node v ∈ Vt at the timestamp si ∈ [St,St+1].**

Two critical factors in dynamic networks are the **proximity between nodes** and the **temporal continuity of nodes over time**:

* **Proximity between nodes** refers to the measure of how closely connected two nodes are within the network at a certain timestamp graph. The proximitycan be influenced by direct connections (e.g. a temporal edge between two nodes) or indirect relationships (e.g., shared temporal neighbours or paths).
* **Temporal continuity of stable nodes** involves maintaining consistent representations of nodes that remain relatively unchanged across time. This is crucial for capturing the evolution of nodes and their roles within the network without introducing unnecessary distortions. For example, consider an article in a specific scientific field that continues to be cited over decades. The node representing this article remains stable in terms of its role in the network, consistently attracting citations as new research builds upon its findings.

## Word2Vec and its inspiration for Dynamic Network Embeddings

Word2Vec is a neural network-based method used to learn continuous vector representations of words by leveraging their context within a corpus. Words that appear in similar contexts are mapped to nearby points in the vector space, capturing semantic relationships between words.

Two key approaches in Word2Vec are:

* **Continuous Bag-of-Words (CBOW):** Predicts a target word based on its surrounding context words.
* **Skip-Gram Model:** Predicts context words given a target word.

These models learn embedding by maximizing the likelihood of observing context words around a target word, using positive and negative sampling to make computation efficient. The proximity of word embeddings in the learned vector space reflects semantic similarity.

Dynamic network embeddings draw direct inspiration from Word2Vec. Instead of modeling words and their contexts, dynamic embeddings model nodes in a network and their temporal contexts:

* **Nodes as Words:** Nodes represent entities (e.g., articles in citation networks), analogous to words in Word2Vec.
* **Node Contexts:** Random walks generate sequences of nodes, capturing the local structure of the network over time. These sequences act as temporal contexts like word co-occurrences.
* **Proximity Learning:** The embeddings preserve node proximity within each snapshot of the dynamic network, like how Word2Vec preserves word similarity.
* **Temporal Evolution:** While Word2Vec assumes a static corpus, dynamic embeddings introduce a temporal dimension to track changes in node relationships and positions over time.

The core idea of learning context-based embeddings to capture relationships is retained in dynamic network embeddings but adapted to represent evolving structures. This adaptation enables dynamic embeddings to model temporal continuity, ensuring that stable nodes maintain consistent representations while allowing for significant embedding drift in trending nodes.

## Citation Dynamic Network Specificities

Citation networks exhibit unique and complex properties that distinguish them from other types of dynamic networks:

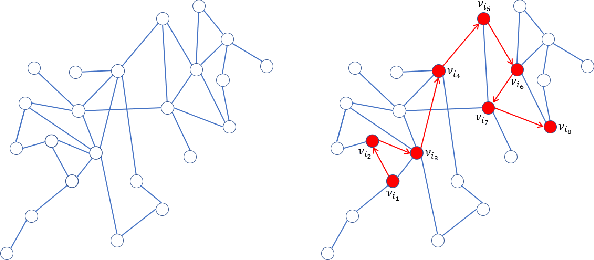
* **Irreversibility of Edges:** A citation, once established, is permanent and unidirectional. This reflects the enduring nature of knowledge transfer, where older foundational works remain integral to the evolving research landscape.
* **Node Longevity:** Articles often remain relevant for years or even decades after publication. This longevity reflects their continued influence as foundational works or seminal contributions in a specific domain. Articles with sustained citation activity serve as anchors in the evolving structure, providing stability in the dynamic embedding space.
* **Citation Burstiness:** Citations are unevenly distributed over time. Articles may experience sudden surges in citations due to emerging trends, paradigm shifts, or groundbreaking discoveries. These bursts are crucial for identifying trending publications and tracking shifts in research focus.
* **Hierarchical Structure of Influence:** Citation networks often exhibit a hierarchical flow of influence, where older, highly cited works act as "roots," and newer works branch out to build upon these foundations. Over time, articles may accumulate more citations, further reinforcing their importance in the network hierarchy.
* **Field-Specific Dynamics:** Different research fields exhibit distinct citation behaviors. For instance, fast-evolving fields such as artificial intelligence (AI) and bioinformatics experience rapid bursts of citations, while fields like mathematics or physics may have longer citation cycles where articles remain influential for extended periods.

Understanding these characteristics is crucial for designing methods tailored to citation network analysis. By accounting for the permanence of edges, the longevity of influential nodes, and the bursty nature of citation events, dynamic embedding methods can better capture the evolving structure and influence of articles within citation networks. These insights allow researchers to identify emerging trends, track the propagation of knowledge, and recognize the enduring impact of foundational works.

## Random Walk

A random walk is a stochastic process that generates a path consisting of successive steps on a mathematical space, such as a graph. At each step, the next node is randomly selected from the neighbours of the current node.

Random walks are widely used for efficiently exploring graph structures, analysing connectivity, and identifying communities. Additionally, they enable effective sampling of large networks, especially when global computations are impractical or computationally expensive.

Basic random walk algorithm on graph:

**Input:**

**Figure 2: Construction sequence of nodes using random walks**

* : A graph with nodes V and edges E.
* : stating node
* : length of the walk
* : Edge weights (optional)

**Output:**

* : A sequence representing the random walk.

**Steps:**

1. Set
2. For to L:
   1. Set
   2. If graph is unweighted:
      1. Select uniformly at random from N
   3. Else:
      1. For each Compute the probably:
      2. Select based on the probability distribution.
   4. Append to
   5. Set as
3. Return

The random walk length (L) is a key parameter that directly influences the behaviour, outcome, and efficiency of graph exploration and the tasks that depend on the random walk, such as graph traversal, connectivity analysis, and node embedding generation:

* **Short walks** in a random walk process focus on exploring the local neighbourhood of a starting node and capturing immediate connectivity. These walks are particularly suitable for tasks requiring fine-grained, localized relationships, such as detecting tightly connected communities or performing micro-level analysis.
* **Long walks** explore a broader region of the graph, capturing global structures and distant connectivity patterns. This makes them useful for understanding macro-level graph properties, such as identifying how nodes are connected across different clusters or detecting bridge nodes that link communities.

Random walks are considered as efficient tools for exploring dynamic networks due to their localized and adaptable nature. They allow real-time updates as networks evolve, enabling tasks like community detection, anomaly detection, node embedding, and influence propagation in changing environments. By leveraging their flexibility and scalability, random walks provide a robust solution for understanding the complex temporal behavior of dynamic networks.

## Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is an optimization algorithm in machine learning and deep learning. It is an iterative method for minimizing an objective function, often the loss function of a model, which is parameterized by a set of weights or parameters.

Understanding the Mechanism of SGD:

1. **Iterative Updates:** SGD works by adjusting model parameters based on how the model performs on individual data points or small subsets of the given dataset.
2. **Random Sampling:** Instead of looking at the entire dataset, SGD uses a randomly selected data point from the dataset at each iteration. This makes it computationally efficient and suitable for large-scale data.
3. **Formula for Update:** At each iteration, the model parameters are updated as follows:

*: Model parameters*

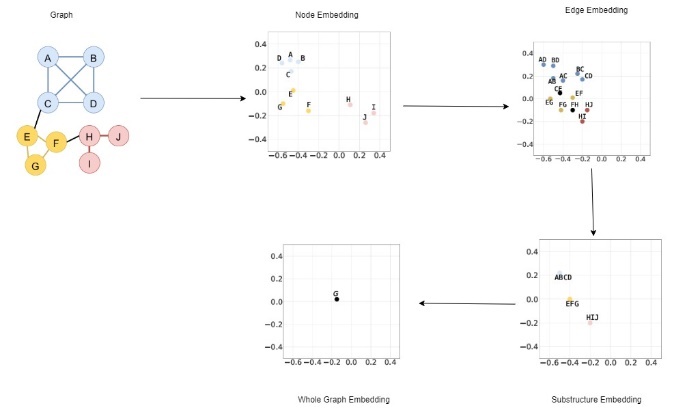
*: Learning rate, controlling the size of the step in the parameter update*

*: Gradient of the loss function computed*

1. **Stochastic Nature:** Due tot the fact that SGD uses random samples, the updates have some stochasticity. This introduces noise into the optimization process, which can help the algorithm escape minimum point values.

By leveraging iterative updates and random sampling, SGD offers a scalable solution for training models on large datasets. While its stochastic nature introduces noise, this can be advantageous in navigating complex optimization landscapes, helping escape local minimum points and improving overall performance.

## Dynamic Bernoulli Embeddings

Dynamic Bernoulli embeddings are a novel approach in dynamic network analysis that aim to preserve the proximity and temporal continuity of nodes in a dynamic network. Unlike methods that treat networks as static structures, this technique incorporates temporal data by embedding nodes into a low-dimensional vector space over discrete time steps. Dynamic Bernoulli embeddings utilize random walks to generate sequences of nodes. This approach ensures that stable nodes maintain continuity across time while capturing the evolution patterns of dynamic nodes.

**Figure 3: A single iteration of dynamic Bernoulli embedding within the same vector space.**

Key Components of the Calculation Process:

1. Embedding Vectors:
   * **Node embedding**  A low-dimensional vector that represents node g at time t. It captures the node’s relationships within the network at that timestamp.
   * **Context vector** A shared vector for each node, representing its role or interactions across all timestamps. It acts as a reference point to determine how nodes relate to their context.
   * **Context Size (:** The number of neighboring nodes considered around a given node in a random walk sequence for learning embeddings. For example, if cs=4, the two nodes before and two nodes after the given initial node in the sequence are part of its context.
2. Node Indicator Vector
   * This is a binary vector where =1 if node g is part of the context for the i-th position in a random walk at time t and otherwise =0.
   * It ensures that the model correctly identifies which nodes are part of a given context.
3. Context Relationship Score :
   * This score measures how well a node g fits its context.

*: Embedding of node g at time t.*

*: Context vector of node g.*

*: Size of the context window.*

* + Calculated as:

1. Likelihood Functions:

To ensure embeddings preserve proximity and temporal continuity, the model defines two likelihood functions:

* **Positive Likelihood** (:

This measures how well the model predicts the true relationships between nodes in the network.

* **Negative Likelihood** (:

Here, represents a sampling distribution for "negative" nodes (nodes not part of the true context). Negative sampling reduces computational complexity.

**Negative Sampling**: Instead of calculating for all possible negative nodes (which would be computationally expensive), a small subset is sampled at random. This speeds up training while retaining accuracy. The number of "negative samples" drawn during the embedding training process is defined as negative sample size (ns).

1. Regularization:

To ensure temporal consistency and prevent overfitting, two regularization terms are added:

* + For the context vector:

This ensures that the context vectors remain stable and bounded.

* + For temporal continuity:

This penalizes large changes in embeddings between consecutive timestamps.

Basic DBE (Dynamic Bernoulli Embedding) algorithm on graph:

**Input:**

* : A low-dimensional vector that represents node g at time t.
* : Current context vector.
* : A sequence of nodes representing the random walk.
* : Nodes group of timestamp t.
* : Context size.
* : Negative sample size.

**Output:**

* Updated values of both and

**Steps:**

1. For each in :
   1. For each in :
      1. Minimize loss by SGD(
      2. Update and
   2. End for
2. End for

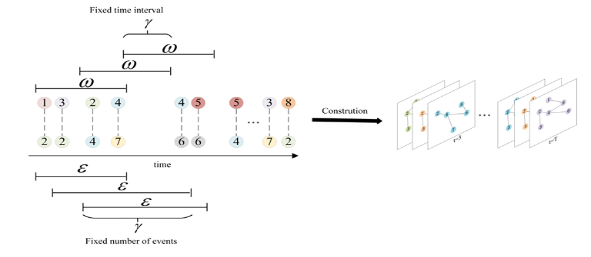
Dynamic Bernoulli Embeddings provide a robust framework for analyzing dynamic networks by preserving both the temporal continuity and proximity of nodes. The model achieves this by leveraging random walks, context-aware embeddings, and regularization techniques. By focusing on computational efficiency through negative sampling and carefully balancing the likelihood and regularization terms, this approach can effectively model the evolution of networks over time.

## Dynamic Network Construction

**Definition 2 (Dynamic Network Embeddings): Given a dynamic network Γ = {G1,...,GT }, dynamic network embeddings aim to project a node д ∈ Vt into a low-dimensional vector space by a mapping function f : д → y(t) | д ∈ RD, D ≪ max|Vt|, t ∈ [1,T].**

Dynamic networks are constructed from timestamped events, such as citations. Each snapshot is derived using either:

1. **Fixed Time Intervals ():**
   * Edges ​ are defined based on events occurring within the time window .
   * This approach ensures temporal consistency but may result in sparse graphs if events are unevenly distributed.
2. **Fixed Number of Events ():**
   * Each snapshot contains a predefined number of events, ensuring uniform density across graphs.
   * However, this approach may break temporal consistency, which can be addressed using overlapping windows.



To ensure smooth transitions and maintain continuity across the dynamic network, an overlap is introduced between adjacent graphs. This overlap creates a connection between consecutive graphs, preserving the flow of information and minimizing disruptions in the network's temporal dynamics.

**Figure 4: Dynamic network construction by the fixed time**

There are four main matrices that represent the Dynamic Network Embeddings:

1. **Embedding Matrix (​):** each row corresponds to a node's low-dimensional vector representation at time . These embeddings are specifically optimized to maintain both the proximity of nodes within the graph and the temporal continuity across different time steps.
2. **Adjacency Matrix (​):** encodes the connections between nodes in the graph at time . In this matrix, indicates the presence of an edge between nodes . This matrix is instrumental in calculating node similarities and serves as a foundation for generating random walks within the graph.
3. **Context Matrix ():** stores context vectors for nodes, capturing neighborhood information derived from random walks. These context vectors are dynamically updated during training to maximize the likelihood of observing true neighboring nodes while minimizing the likelihood of random connections through negative sampling.
4. **Temporal Regularization Matrix:** designed to align the embedding matrices and by penalizing significant differences in the embeddings of stable nodes across consecutive timesteps. This mechanism ensures smooth temporal transitions in node representations, maintaining consistency and stability in the embeddings over time.

The embedding process can be summarized as follows:

* **Input:** A dynamic network , random walk parameters , embedding size (), context size (), and negative sampling size ().
* **Process:** For each snapshot , random walks generate sequences of nodes, which are used to update embedding and context matrices (​ and ).
  + **Output:** Embedding matrices ​ that represent the temporal and structural properties of the network.

A screenshot of a computer

Description automatically generated

**Figure 5: Dynamic Network Embedding model structure**

# Approach

## Tracking Evolution Articles in Citation Networks

To detect a trending article in a dynamic citation network using the Dynamic Network Embedding approach, we analyze how the node embeddings (representing articles) evolve over time. A trending article exhibits distinct and measurable behaviors that reflect its increasing importance and impact within the network:

* **Rapid Increase in Degree:** A trending article receives a sudden and notable rise in citations (edges) within a short time window. This rapid increase signifies growing recognition of the article's relevance and its influence on subsequent research. Such articles often address emerging topics, present groundbreaking methods, or offer significant insights that capture the attention of the academic community.
* **Significant Embedding Drift:** The position of the article in the embedding space changes substantially between consecutive timesteps. This drift occurs as the article attracts citations from diverse or unexpected areas, signaling its expansion beyond its original domain. Embedding drift highlights the article's evolving role and its capacity to influence interdisciplinary research or spawn new lines of inquiry.
* **Centrality Growth:** Over time, the article becomes more central within the citation network. Centrality measures, such as betweenness, closeness, or eigenvector centrality, indicate how influential a node is within the network. An article experiencing centrality growth acts as a "hub" connecting multiple research areas, fostering collaborations, and serving as a foundation for subsequent work.

These behaviors collectively provide a comprehensive view of an article's trajectory within the evolving citation network. By tracking degree growth, embedding drift, and centrality changes, researchers can identify not only articles gaining influence but also emerging research themes and fields. Combining these metrics offers a robust, multi-dimensional method for detecting trending articles and understanding their evolving impact on the research landscape.

# Testing Plan

# References