Open-Source AI Community as "Trading Zone": The Role of Open-Source Models in the Diffusion of Artificial Intelligence Innovation

Completed Research Paper

Kaige Gao
Case Western Reserve University
Cleveland, OH, USA
kxg297@case.edu

Youngjin Yoo
Case Western Reserve University
Cleveland, OH, USA
yxy23@case.edu

Aaron Schecter
University of Georgia
Athens, GA, USA
aschecter@uga.edu

Abstract

Previous studies have highlighted the role of computational models in the diffusion of AI research, suggesting that the impact of models extends beyond their mere existence. In our analysis, which encompasses 64,059 papers and 98,105 models from 2000 to 2023, we adopt a twofold approach, integrating Graph Neural Network (GNN) and traditional network analysis to provide complementary insights. We discovered that a model's dependency network significantly influences paper citations and warrants further attention. Notably, the placement of focal versus offspring models in relation to a paper markedly affects its diffusion. Our findings enrich the literature on AI innovation diffusion by underscoring the critical importance of models and their networks. Additionally, our work enhances the GNN literature by advancing the capabilities of existing explainers to accommodate the nuances of temporal network structures

Keywords: Network Analysis, Evolution of AI Innovation, Diffusion of AI Innovation

Introduction

Artificial Intelligence (AI) has become a hot topic, attracting widespread attention due to its rapid diffusion and application across various domains (Davenport and Ronanki 2018; Makridakis 2017). This swift and extensive adoption has sparked significant interest in understanding the factors that influence the diffusion patterns of AI innovations (Baek et al. 2020; Fredström et al. 2021).

In their exploration of AI's proliferation, researchers have utilized various proxies to study its dissemination. These include AI-related job postings (Acemoglu et al. 2022; Alekseeva et al. 2021), patent filings (Baek et al. 2020; Fredström et al. 2021), and especially, scholarly publications (Frank et al. 2019; Tang et al. 2020). Such publications, authored by both academics and industry professionals, are pivotal to the propagation of AI innovations, representing a predominant channel for the exchange of research insights (Tang et al. 2020). An analysis of the dissemination and traction of AI research is therefore instrumental in discerning the overarching trends in the distribution of AI advancements.

Previous research on the citation and diffusion of scholarly work has often centered on the influence of authors' social networks and the intrinsic attributes of the publications on their subsequent popularity and impact. For example, Uzzi et al. (2013) discovered a greater propensity for papers that strike a balance between novelty and conventionality to achieve popularity and exert a significant impact. White (2001) demonstrated a tendency for researchers to cite publications authored by individuals within their own social networks, suggesting that the social dimension plays a crucial role in citation behavior within academic communities.

Emerging studies have turned the spotlight on the contribution of executable models to the circulation of academic papers, with a particular focus on the citation effects of models made public alongside the research (Bhattarai et al. 2022; Kang et al. 2023). Yet, the relevance of these models extends beyond mere availability. They embody a combinatorial evolutionary process, evolving and integrating upon existing models to create a dynamic, interdependent ecosystem. This process is not confined to academia's citation networks but also thrives within the open-source community, reflecting an intricate interplay that influences academic trajectories and innovation waves alike (Boland Jr et al. 2007).

This study probes deeper into the effects of model dependency networks on the dissemination of AI research. Through a composite dataset drawn from Papers with Code, Open Alex, Semantic Scholar, and GitHub, featuring 64,059 papers and 98,105 models from 2000 to 2023, we pursue a twofold analysis. Firstly, we employed a Heterogeneous Temporal Graph Neural Network to elucidate the overarching impact of model dependency networks. Secondly, we analyzed the influence of a model's position within the model dependency network on the popularity of research.

Our findings reveal that models hold comparable—if not greater—significance than authors in swaying a paper's scholarly impact, meriting more focused academic attention. Notably, papers whose focal models at the nexus of these networks often see enhanced citation frequencies. In contrast, papers linked to a multitude of centrally located yet offspring models might encounter a diffusion of focus. This research enriches the diffusion of innovation discourse by shedding light on the understudied role of computational models in determining the reach of AI research. It also advances the field of graph neural network analysis by augmenting the capabilities of existing explainers to accommodate temporal network complexities.

Literature Review

Previous Study on Diffusion of Papers

In the realm of artificial intelligence (AI) research, the publication of academic papers is a primary vehicle for the dissemination of new innovations. Articles introducing core AI developments serve as the conduit for knowledge sharing within the scholarly community. For instance, the paper "Language Models are Few-Shot Learners" by Brown et al. (2020) heralded the arrival of OpenAI's GPT-3 and established the groundwork for subsequent iterations like ChatGPT. Such contributions encapsulate detailed explorations of innovations, stimulating further analysis, development, and adoption by academics and practitioners, thus propelling the spread of innovative ideas as outlined by Rogers (2010).

Continuing on this trajectory, the building blocks of AI advancements are often embedded within preceding publications, with newer works frequently citing foundational studies, as conceptualized by Arthur (2009). This interconnectivity not only advances the field but also offers a lens through which to observe the momentum and direction of AI innovation.

Prior research has illuminated various elements that contribute to the reach and assimilation of academic writings. On the one hand, paper-centric attributes—like intrinsic quality, initial recognition, domain specificity, and structural characteristics such as length and the inclusion of visual aids—have a pronounced impact on the paper's future citations and its prominence within the field (Bornmann and Daniel 2008). Uzzi et al. (2013) provide evidence that a balance between novelty and traditionalism can enhance a paper's impact and reception. On the other hand, author-centric factors, including an author's standing in the academic community, the presence of funding, and the breadth of their social networks, also significantly sway a paper's diffusion (Fleming et al. 2007; Uzzi and Spiro 2005). Notably, Fleming et al. (2007) suggested that authors who exhibit generative creativity and engage in social network brokerage tend to produce more influential research. The role of models related to a paper, while being ignored, is starting to get scholars' attention.

The Role of Models on the Propagation of AI Research

In the AI domain, models are seminal, functioning not merely as vessels for housing algorithms but also as conduits that connect expansive datasets with practical applications. Models hold a multifaceted role: they can be pioneering as standalone breakthroughs (Gao et al. 2021), integral as part of larger systems (Silver et al. 2016), or foundational in the conceptualization of new technologies (Vaswani et al. 2017). Their intrinsic value lies not only in their technical construction but also in their influence on the scientific community's reception of AI research~~.~~

The presence of public models enhances research transparency and facilitates the replication and validation of scientific work (Mueller-Langer et al. 2019). Innovation in AI occurs both in academic settings and within the open-source community, the latter becoming a crucial platform for AI development. To integrate resources from these domains effectively, scholars are making concerted efforts to establish connections between them. For instance, Shao et al. (2020) introduced 'paper2repo', a recommendation system that identifies relevant GitHub repositories based on academic papers. Similarly, initiatives like Papers With Code (Stojnic et al. 2022)) and platforms such as RedditSOTA demonstrate efforts to maintain traceability between academic papers and corresponding models, promoting a well-organized open science environment.

This flourishing open science ecosystem contributes to the proliferation of AI innovations. Scholars are increasingly examining how public repositories influence the diffusion of AI papers. Previous research has shown that papers with publicly available code are likely to receive higher citation counts (Bhattarai et al. 2022; Bonneel et al. 2020) or experience a significant increase in citation rates (Kang et al. 2023).

However, the significance of models extends beyond their mere existence. Within the open-source community, models evolve based on dependencies from previously released packages, forming an independent ecosystem that parallels the academic sphere. Conversely, academic progress in papers evolves through building upon existing knowledge, often cited from earlier works. These two streams occasionally intersect in a 'trading zone', which may lead to new patterns of AI innovation like wakes of innovation (Boland Jr et al. 2007). Understanding these dynamics could highlight the theoretical importance of network effects in scientific innovation and suggest a broader model of 'combinatorial evolution' in technology and science. This interaction often leads to significant shifts in both fields, creating a symbiotic relationship where academic insights and open-source developments inform and enhance each other, contributing to the broader landscape of technological evolution. Therefore, it is essential not to focus solely on the presence of models but to also consider the dependency relationships within model networks and their impact on the diffusion of AI papers.

Network Perspective on Citation Analysis

Network analysis is a fundamental approach in citation analysis due to its natural alignment with the inherent structure and dynamics of citation data. This methodology effectively captures the flow of knowledge and influence between papers, allowing researchers to visualize and quantify these relationships. It reveals key patterns such as central papers or authors, the spread of ideas, and the structure of academic communities.

Researchers using network analysis for citation studies typically employ two main methodologies: traditional statistical methods and network embedding. Traditional statistical methods involve extracting network metrics such as centrality (Yan and Ding 2009; Zingg et al. 2020), clustering coefficients (Li et al. 2007; Ren et al. 2014), and community structures (Chen 2012; Takeda and Kajikawa 2010), then integrating these metrics into regression models to analyze their effects on citation dynamics. For instance, Yan and Ding (2009) utilized degree centrality, betweenness centrality, and PageRank from a coauthor network to explore the impact of author influence on paper citation counts. Similarly, Chen (2012) analyzed structural variations through metrics like modularity change rate and centrality divergence to predict their effects on citation counts. This method is effective in quantifying the influence of various network characteristics and provides robust explanations of mechanisms influencing citation behaviors.

Network Embedding represents a more recent methodology that transforms the citation network into a lower-dimensional vector space (Tang et al. 2015). The resultant vectors capture the network's structural nuances, which can then be utilized in various machine-learning algorithms to predict outcomes (Hou et al. 2020). This approach has been widely adopted to generate node features that encapsulate both the direct and indirect structural information within the network. For example, Choi and Yoon (2022) utilized network embedding to transform technical elements into vector representations, which enabled them to measure distances within the vectors as a means of exploring patent knowledge in citation networks. While the network embedding-based approach has been adopted to generate node features, it typically does not incorporate company demographic information. Therefore, scholars have expanded their methodologies to include advanced techniques such as Graph Neural Networks (GNNs), which integrate network embeddings with deep learning to predict complex network behaviors directly from the graph's structure, thereby maintaining and leveraging network topology throughout the learning process (Zhou et al. 2020). For instance, Holm et al. (2020) utilized a temporal GNN framework to predict citation counts over years, demonstrating superior performance over standard LSTM and graph convolutional neural networks. Additionally, Cummings and Nassar (2020) employed a GNN-based architecture to predict papers' citation counts at the time of publication, circumventing the need for initial citation trends. Besides delivering good performance, GNNs also benefit from GNN explainers, which provide post-hoc explanations (Ying et al. 2019a). These explainers can demystify Blackbox operations of GNNs by identifying which subgraphs or features are most influential in specific decisions made by the model (Ying et al. 2019b). GNN explainers have been widely used in the fields of chemistry and biology to aid in molecular property prediction (Wu et al. 2023) and drug discovery (Wang et al. 2023). However, they do have some limitations, and some researchers have started to express concerns about their reliability and robustness (Agarwal et al. 2023; Zhang et al. 2024a).

Therefore, in this research, we propose a two-part analysis to utilize both traditional statistical methods and GNNs to complement each other, investigating the influence of models on the diffusion of AI papers.

Firstly, we apply the Graph Neural Network framework to broadly examine the impact of model dependency networks on the citation counts of papers. The application of GNNs is particularly suitable for this analysis, as they adeptly capture a wealth of information without diminishing data integrity (Xu et al. 2018). Unlike traditional methods that may condense or overlook complex network structures, GNNs consider the full topological data structure, which can potentially reveal insights about the collective influence of models that traditional statistical analyses might miss, especially in cases where the network's topology plays a critical role in the diffusion process. Through this study, we seek to affirm the hypothesis that not merely the existence of open-source models but also the dependency relationships among these models significantly influence the popularity of AI research.

Secondly, we implement statistical models to uncover the detailed relationships between models and papers, thereby complementing the overarching insights provided by GNNs and ensuring the robustness of previous findings. While GNNs yield importance scores for models and their features, they may not effectively quantify the exact effects on citation counts. Our statistical model will probe these details, enabling us to understand the nuanced mechanisms of the model network's effect on the dissemination and recognition of AI research.

Data

Our analysis integrates data from four distinct sources: (1) Papers with Code (http://paperswithcode.com/), (2) OpenAlex, (3) Semantic Scholar, and (4) GitHub. Figure 1 summarizes the data process and analysis structure.

Papers with Code

Following Kang et al. (2023), we utilize Papers with Code (PwC) as our primary database. Designed by the Meta AI team and officially partnered with arXiv since October 2020, PwC is the largest platform linking ML research articles with their corresponding repositories. It has been widely used in research to build knowledge graphs between papers and code (Kannan et al. 2020), analyze the impact of public code availability on citation counts (Kang et al. 2023), and summarize code availability in top AI papers (Lin et al. 2022).

Our research focuses on analyzing records from PwC, specifically those papers published between January 2000 and December 2022. During this time span, the PwC dataset comprises 335,354 papers, of which 168,497 are associated with open-source repositories containing models. Our study primarily targets papers with associated models because Kang et al. (2023) demonstrated that papers with available open-source model experience a 20% increase in monthly citations compared to those without. Building on previous study, this research aims to explore how the model dependency network influences paper citations beyond models' mere existence.

We restrict our analysis to papers categorized under the machine learning portal, excluding those related to astronomy, physics, computer science, statistics, and mathematics. We also exclude 884 models not from the GitHub community, focusing on the dependency relationship inside the GitHub community. Our final data set from Papers with Code contains 76,524 papers and 98,703 models, along with 127,366 paper-model relationships.

OpenAlex

We use OpenAlex (https://openalex.org/) to collect related paper information and author information. We chose OpenAlex for two reasons: first, it is a fully open platform integrating multiple data sources, including Microsoft Academic Graph (MAG), ORCID, Crossref, and Unpaywall, and it has been widely used in academic research (Priem et al. 2022; Schares and Mierz 2023) ; second, Microsoft Academic Graph was discontinued on December 31, 2021, which does not match our main data source time span from PwC and researchers have found OpenAlex at least as suited for bibliometric analysis as MAG was (Scheidsteger and Haunschild 2023).

However, we note that OpenAlex records the creation date for each paper, the date it started documenting the paper's information. This can cause a problem: citation counts a paper gained before the documentation date will not be captured. For example, the paper "YOLOv3: An Incremental Improvement" published in 2018 had its record created in 2022 by OpenAlex, causing citation records from 2019 to 2022 to be missing. Therefore, we filtered out those documents whose publication date and the documentation date of OpenAlex differ by more than one year, removing 2,544 papers. Additionally, 6,391 papers were excluded due to mismatches between the Papers with Code and OpenAlex datasets. Moreover, due to significant missing data regarding author institutions reported by OpenAlex (Zhang et al. 2024b), we chose not to include this information in our analysis. However, we did include detailed author-related information such as citation counts over years, number of works published over the years, research topics, h-index, and second-year citedness in our study.

Semantic Scholar

While cleaning the reference information from OpenAlex, we found that it did not fully match with the original papers. After a manual random check, we determined that the reference lists provided by Semantic Scholar are basically accurate. Launched by the Allen Institute for Artificial Intelligence in 2015, Semantic Scholar has established a significant presence in the academic community and is recognized for its advanced AI and machine learning technologies applied to the scientific paper search and analysis (Fricke 2018). Its dataset has been widely used to build literature graphs (Ammar et al. 2018; Kinney et al. 2023). While it may not provide as extensive information as OpenAlex, we decided to use Semantic Scholar as a complement to OpenAlex on the reference list. We collected the reference list of each paper in our dataset and constructed internal citation relationships. Following other widely used citation datasets like Cora and CiteSeer (Cabanes et al. 2012; Giles et al. 1998), we extracted citations only among the papers within our predefined set and ignored any citations referencing papers outside of this set.

GitHub

Since 99.2% of the models connecting with papers from Papers with Code in our dataset are from the GitHub platform, and GitHub is the largest platform hosting open-source models (Cosentino et al. 2017), we collected model dependency graphs and related model meta-information, such as descriptions, topics, languages, and README files from GitHub to construct the model dependency network. After integrating the four datasets, we conducted a cleanup process to ensure the quality and relevance of our data. This involved removing approximately 2,000 papers that lacked corresponding records in Semantic Scholar and about 500 models from GitHub, which were no longer accessible (resulting in 404 errors). Following these adjustments, our final dataset comprises 64,059 papers and 98,105 models.

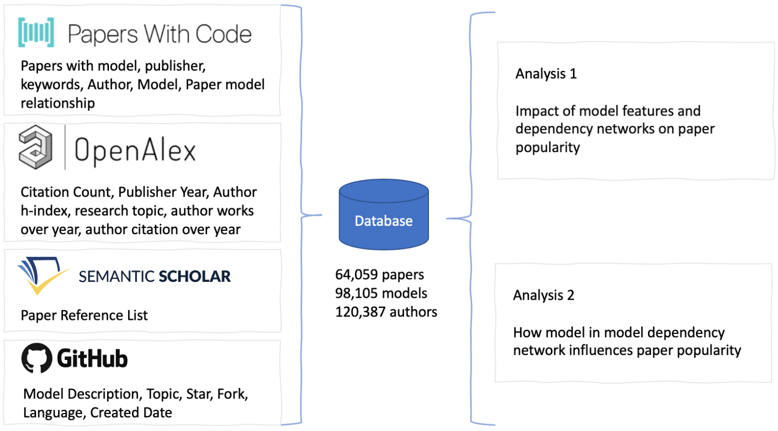
Papers With Code
Papers with model, publisher, keywords, Author, Model, Paper model relationship
OpenAlex
Citation Count, Publisher Year, Author h-index, research topic, author works over year, author citation over year
SEMANTIC SCHOLAR
Paper Reference List
GitHub
Model Description, Topic, Star, Fork, Language, Created Date
Database
64,059 papers
98,105 models
120,387 authors
Analysis 1
Impact of model features and dependency networks on paper popularity
Analysis 2
How model in model dependency network influences paper popularity

Figure 1. Data Processing and Analysis Overview

Analysis 1: Overall impact of model dependency network via GNN

In our first analysis, we aim to demonstrate that factors beyond the mere existence of models, such as models' dependency relationships, model-paper relationships, and models' characteristics, significantly influence the increase in paper citations. Therefore, we place particular emphasis on the network structure and the graph's demographic characteristics, employing a Graph Neural Network (GNN) framework that preserves the integrity of the network structure.

Given the dynamic nature of the relationships among three distinct types of nodes—authors, papers, and models—over time, we constructed a heterogeneous temporal network to capture these evolving interactions. Due to the variability in paper keywords, author research topics, and model descriptions, we embedded these textual elements separately as node features for all node types.

Our model stabilized, converging to a consistent Mean Absolute Error (MAE) of approximately 4.07. The variability in our dataset's citation counts ranges widely, from 0 to a maximum of 255,438 annually, with a variance of 13,163.25. Given this broad range, we consider this level of precision to be satisfactory for the purposes of our study. Furthermore, we have enhanced the existing heterogeneous graph neural network (HGNN) explainer (Ying et al. 2019b) to accommodate temporal networks. This advancement enables the explainer to consider multiple temporal graphs and provide both edge and feature importance for post-hoc analysis, thereby facilitating the identification of critically influential model-related edges and features

Heterogeneous Temporal Network Construction

A heterogeneous temporal network is a graph that encapsulates multiple types of entities (nodes) and relationships (edges) across various timestamps (Fan et al. 2022). In our research, which spans multiple temporal segments, a heterogeneous temporal network is essential for preserving the intricate details of relationships such as author-to-paper, paper-to-paper, and paper-to-model interactions. This structure is particularly advantageous for encapsulating the multifaceted dynamics and influences across the scholarly domain. Given that the impact of a paper accumulates over time and focusing solely on papers published in a specific year and their associated data would result in a very sparse network, our network aggregates data up to the current year, '*t*', which results in a more robust and interconnected dataset.

Previous studies have highlighted the significance of various attributes related to authors, publishers, and papers (Bornmann and Daniel 2008; Uzzi and Spiro 2005). In constructing our network, we have endeavored to incorporate a rich set of features proven to be influential in the diffusion of scholarly work. Our network comprises three primary types of nodes—papers, authors, and models—each embedded with features tailored to their respective roles:

- Author Nodes: Features such as h-index, i10-index, and 2-year mean citation count are used to represent an author's influence and reputation. Additionally, research interests are captured using word2vec embeddings of relevant keywords.
- Model Nodes: Features like the number of GitHub stars and forks, programming language, and textual descriptions are embedded using word2vec to reflect the technical and community engagement aspects.
- Paper Nodes: Information such as publisher details, the number of countries from which authors originate, and the diversity of institutions are included alongside keyword-based features, again embedded using word2vec.

Our network includes four types of edges: paper linking model, paper citing paper, model depending on model, and paper written by author. This deliberate inclusion of diverse attributes ensures that even when the analysis is controlled for various known influences, the significance of models in enhancing scholarly diffusion remains demonstrably evident. Figure 2 presents an example of our heterogeneous temporal graph.

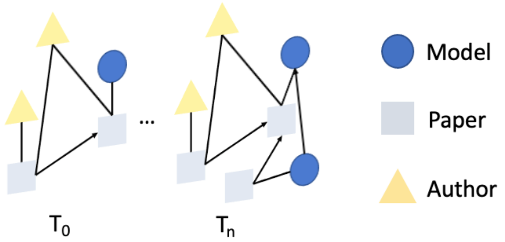

Figure 2. Example of Heterogeneous Temporal Network

GNN framework and Result Explanation

After constructing the heterogeneous temporal network, we proceed to embed both temporal features and edge attributes before inputting them into heterogeneous temporal graph neural network model (Fan et al. 2022). The architecture of our model, referred to as HeterogeneousTemporalCitationGNN (HTCGNN), integrates these embeddings through multiple layers, culminating in a set of forward layers specifically designed to aggregate and interpret edge and feature importance scores. This design enables us to isolate and understand the factors most significantly impacting paper citation rates. We use Mean Absolute Error (MAE) as our loss function because our prediction is citation count, MAE can quantify the average magnitude of errors in a set of predictions, without considering their direction. Contrary to existing approaches like the heterogeneous temporal GNN explainer (Li et al. 2023), which focuses solely on edge importance within sample graphs and overlooks node features, our method incorporates a comprehensive view of both node and edge attributes across the full graph. This broader approach allows for a more detailed explanation of model predictions.

Figure 3 illustrates the overall framework of this process, showcasing how the constructed network feeds into the HTGNN model, which then utilizes the explainer module to derive and aggregate importance scores for both edges and node features.

To systematically assess the overall importance of the model network, we designed and compared four distinct models:

1) Baseline: Comprises only papers and their citation networks, serving as the foundation for comparative analysis.
2) Add Author: Extends the baseline by incorporating relationships between authors and papers, enabling an examination of the influence of scholarly collaboration.
3) Add Model: Builds upon the baseline by adding model nodes and their dependencies, which allows us to evaluate the impact of model interactions on scholarly output.

4) Full: Integrates papers, authors, and models, along with all respective relationships, providing a comprehensive view of the academic landscape.

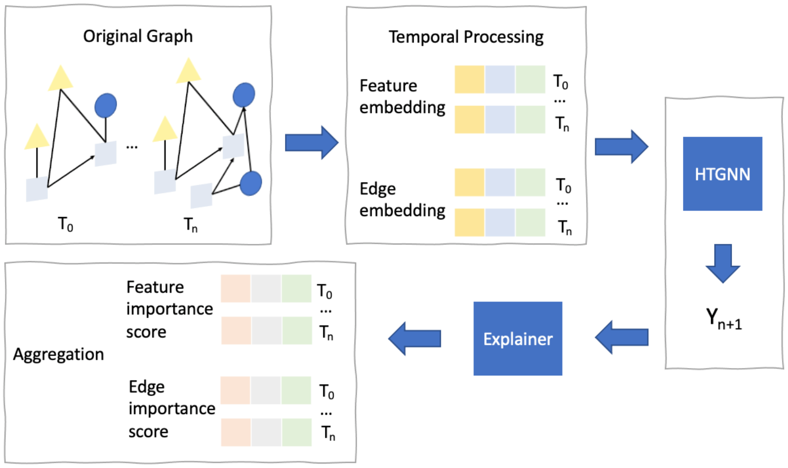

Figure 3. Framework of HeterogeneousTemporalCitationGNN

Similar to the R-squared metric, explained variance score (EVS) is commonly used to evaluate a model's performance by quantifying how well it accounts for the variability in the observed data (LaHuis et al. 2014). It is calculated using the formula $EVS = 1 - \frac{Var(y-\hat{y})}{Var(y)}$, where y represents the observed data and $\hat{y}$ represent the predicted value. We compared the explained variance across these four models to gauge the incremental value added by including different network components.

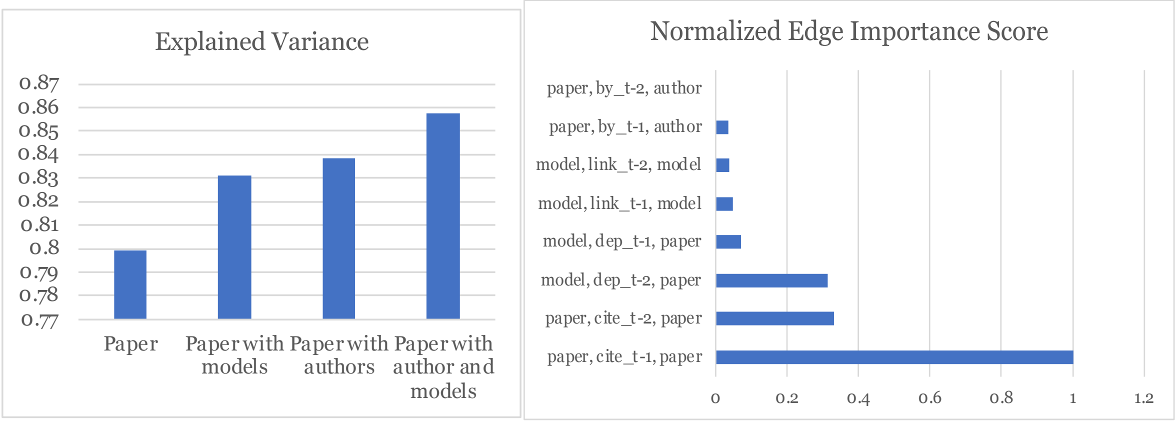

Figure 4. Explained Variance Comparison and Normalized Edge Importance Score

As shown in the left graph of Figure 4, our analysis revealed that the inclusion of models and their dependencies enhanced the explained variance by over 3% compared to the baseline model. Conversely, excluding these elements from the full model led to a 2% decrease in explained variance, underscoring their significance in the network. Due to computational limitations associated with processing large-scale networks, we restricted our analysis to the past two-time spans. This approach addresses memory constraints without significantly compromising the breadth of temporal insights, as the most recent interactions are often the most influential in dynamic academic networks. Figure 4 (right graph) illustrates the importance of various types of edges within our network for the time frames t-2 and t-1, predicting citation counts for papers in year t. This visualization highlights that the structure of the paper's citation network in recent years is the most critical factor influencing citation counts, followed by the model-paper connection. The model-model dependency relationships also play a significant role, albeit less than the direct model-paper interactions. The least impact is observed from the paper-author relationships.

We also calculated the feature importance scores, which were normalized and exclude word embedding features due to their complex interpretability. Figure 5 presents these normalized feature importance scores. The comprehensive plot that includes word embedding features for the top 50 important features has been omitted due to space limitations. Figure 5 reveals that among the explainable features, the number of stars of models holds the highest importance, followed by the influence of authors and the reputation of publishers. These findings underscore the pivotal role of model features, such as the number of stars and forks, and highlight the significant impact of model performance and recognition from developers in the open-source community on the dissemination of papers within the academic field.

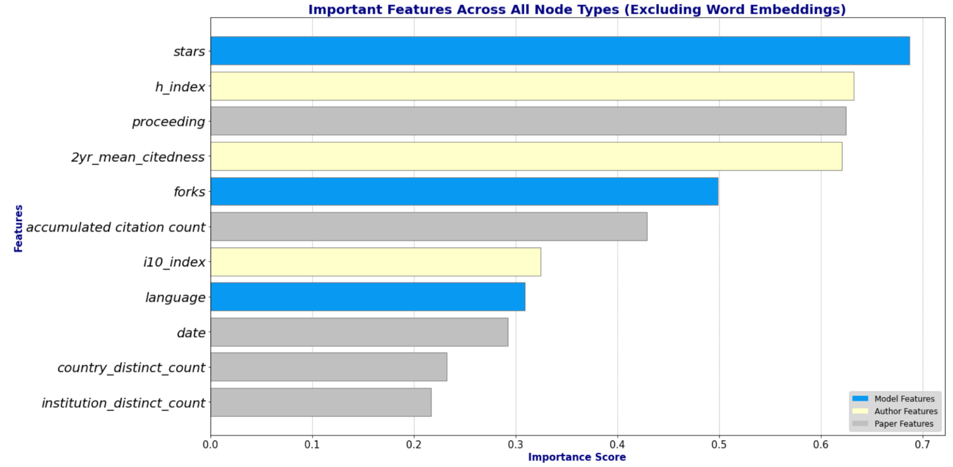

Figure 5. Normalized Feature Importance Score

The analysis demonstrates that the overall importance of models is on par with that of authors in papers related to the AI field. While the influence of authors on paper citations is well-established (Nanumyan et al. 2020), the impact of models remains less explored. Previous studies have shown that AI papers with publicly available code and models tend to receive higher citation counts compared to those without (Kang et al. 2023). Building on these findings, our study reveals that the mere presence or absence of a model is not the only factor of importance; the model networks and open-source community serve as a reflective environment that influences the academic sphere. The relationships among models and their performance within these networks also significantly affect the diffusion and popularity of papers in academic fields.

In our second analysis, we will delve deeper into the specific relationships between models the popularity of related papers. Specifically, we aim to explore how the position of models within the model dependency network affects future citation counts of associated papers.

Analysis 2: Effect of Model Location on Paper Popularity

In our second analysis, we aim to explore the distinct roles that models play in the diffusion of AI research papers. The AI model network consists of two types of models: focal model and offspring model[1]. When papers are published, they may also release associated models on platforms like GitHub. This creates a direct and officially recognized link between the paper and the model. We refer to these models as focal models. Alternatively, other developers may be inspired by the ideas presented in the papers and create their own models based on these concepts. These models are not formally acknowledged within the paper's methodology section and might only be mentioned in supplementary materials or identified through data or code repositories. These implementations might be contributed by other researchers or enthusiasts who have read the paper and decided to code the described models or methods themselves. We refer to these as offspring models.

Both types of models and the model they build on form what we call the 'model dependency network.' Typically, a research paper will have one "focal" code repository, which is the one provided or endorsed by the original authors of the paper, and it can have multiple offspring codes. This network illustrates how models can be both derivative of and foundational to other models within the open-source community. In this analysis, we focus on these dynamics within the model dependency network, seeking to answer the question: how do focal and offspring models' locations influence its related paper's diffusion?

Hypothesis and Theory Development

Models located in central positions within the model dependency network are likely to be more visible and accessible to a larger portion of the network. This increased visibility can lead to higher citation counts as more researchers encounter and utilize these models in their work. According to Pósfai and Barabási (2016), in a network, new entities usually tend to connect to well-established entities, reinforcing their prominence. This dynamic can result in a model that occupies a central network position and possesses high 'prestige' being more frequently adopted in new research, thereby attracting citations to related papers. This phenomenon is supported by the concept of cumulative advantage and the Matthew effect, where early recognition generates further recognition, creating a self-reinforcing cycle (Petersen et al. 2011).

Models in central positions are often stable, fundamental, and well-integrated with other tools and frameworks, facilitating their adoption and application (Chung et al. 2012). This usability reduces the barriers to entry for researchers using these models in new research, which can lead to increased citations (Bonaccorsi and Rossi 2003; Brynjolfsson and Kemerer 1996). Focal models typically carry a stamp of approval from either the original authors or from reputed organizations, enhancing their perceived reliability and authority. When these models also occupy central positions in the model dependency network, they become hubs of innovation and trusted sources, thereby attracting more citations. Thus, we have the following hypothesis:

Hypothesis 1: Papers with focal models that have a more central position in the model dependency network will experience a higher increase in citation count in the following year

offspring models, although not endorsed by the original authors, often emerge from the community and can incorporate novel approaches and techniques absent in focal releases. Their central position in the model dependency network suggests these innovations are recognized and utilized by other developers and researchers, partially reflect the popularity of the ideas contained in related papers, boosting the citations of the papers associated with them, as researchers may trace back to the original papers for deeper insights.

Central offspring models benefit from active community engagement; the feedback and improvements made by a diverse group of developers can enhance the model's robustness and applicability. This active community validation often leads to increased citations as the community vets and refines the model (Bhattarai et al. 2022). Furthermore, offspring models, which can be considered as informal intellectual collaboration with central topics, often fill gaps or specific needs that focal models do not address. Their centrality indicates they play critical roles in the network, perhaps offering functionality that complements

[1] In the Papers with Code dataset, the focal model refers to the official model, while the offspring model refers to the unofficial model.

focal models, which can attract citations from researchers seeking specific solutions. The central role of offspring models in a dependency network also facilitates the cross-pollination of ideas between various fields and disciplines, leading to a broader impact and higher citation counts as the model influences a diverse range of academic areas (Neville and Jensen 2007). Therefore, we have the following hypothesis:

Hypothesis 2: Papers with offspring models that have a more central position in the model dependency network will experience a higher increase in citation count in the following year

Variables

We use the annual increase in citation counts of papers (denoted as $Cite_{i,t}$, which represents the citation count increase for paper i in year t) as our dependent variable, which represents the popularity of AI research over years.

Given that the model dependency network is a weighted directed network, we use in-degree eigenvector centrality to estimate the influence or positioning of paper-related models within this network. In-degree eigenvector centrality is particularly apt for this analysis as it not only measures the direct dependencies of a model but also assesses the quality of these connections by considering the influence of the connecting nodes (Bonacich 2007).

We calculate the in-degree eigenvector centrality for both focal and offspring models separately, allowing us to explore potential differences in their impacts. Furthermore, considering the frequent occurrence of multiple offspring models associated with a single paper, we evaluate both the average centrality of offspring models and the centrality of the most influential offspring model. This dual approach helps us determine whether it is the collective impact of all associated models or the influence of the most prominent model that affects a paper’s popularity.

We account for the effect of last year's citation count to control for potential momentum, acknowledging that a paper's previous citations can influence its future citations (Price 1976). We also control for the out-degree eigenvector centrality of models, which reflects both the innovation potential and stability of a model. A high number of dependent relationships (out-degree) might indicate less stability, as frequent changes or updates may be needed to maintain relevance. Conversely, such dependency could also suggest limited scope for groundbreaking innovations, as the model primarily influences existing developments. This metric not only considers the quantity of these outward connections but also their quality, particularly how influential or central the dependent models are within the network. We analyze this separately for focal and offspring models.

To control the influence of authors on paper citations, we include both the average and maximum outputs related to their scholarly activities. Specifically, we include the average and maximum number of works published per year by authors, as well as the average and maximum increase in citation counts their works receive annually. Including both average and maximum values help to account for varying levels of productivity and influence among authors. The average metrics provide insights into the typical performance and influence of the author group, which is particularly useful in settings with collaborative research such as labs with multiple contributors. In contrast, maximum values are included to capture the potential outsized impact of highly prolific or influential individual authors within a group, recognizing that a single high-performing author can significantly enhance a paper's visibility and citation potential. By using these combined metrics, we aim to balance the overall contribution of the author group against exceptional individual contributions, ensuring a more nuanced understanding of author impact on paper popularity.

We also control for unobserved heterogeneity and time-related variability by incorporating fixed effects for papers and years. This approach accounts for factors that remain constant over time for each paper, such as the publisher, the number of authors, and keywords. These attributes are effectively captured in the paper fixed effects, justifying their omission from the variable set explicitly modeled.

Given the skewed distribution of in-degree and out-degree eigenvector centrality measures, we implemented a logarithmic transformation on these variables to mitigate the skewness. Subsequently, we standardized the transformed data to achieve a normal distribution, enhancing the robustness of our statistical analyses. The table below details the distribution of all variables post-transformation that are included in our study.

	Mean	SD	Min	Max	Median
Citation Increase (1)	17.175	152.161	0.000	25438.000	2.000
Focal Model Location (2)	0.000	1.000	-0.515	4.069	-0.515
Focal Model Dependency (3)	0.000	1.000	-0.061	23.127	-0.061
Offspring Model Max Dependency (4)	0.000	1.000	-0.106	11.526	-0.106
Offspring Model Most Central Location (5)	0.000	1.000	-0.610	3.313	-0.610
Offspring Model Average Dependency (6)	0.000	1.000	-0.106	12.645	-0.106
Offspring Model Average Location (7)	0.000	1.000	-0.606	3.516	-0.606
Last Year Citation (8)	0.000	1.000	-0.082	251.045	-0.075
Average Number of Authors Published Work (8)	0.000	1.000	-0.399	37.336	-0.279
Average Citations of Authors (10)	0.000	1.000	-0.450	29.713	-0.319
Max Number of Authors Published Work (11)	0.000	1.000	-0.359	42.679	-0.297
Max Citations of Authors (12)	0.000	1.000	-0.409	9.049	-0.328

Table 1. Description of Variables

Model and Result

To estimate the impact of a model's centrality on the subsequent year's citation count increase of related papers, we employ a generalized Poisson fixed effect model. This approach allows us to handle count data with overdispersion and to assess fixed effects accurately (Consul and Famoye 1992; Saputro et al. 2021), The specification of our model is as follows:

$$\boldsymbol{Cite_{i,t} == \beta_0 + \beta_1 * FocalModelLocation_{i,t-1} + \beta_2 * OffspringModelMostCentralLocation_{i,t-1} + \beta_3 * OffspringModelAverageLocation_{i,t-1} + Control\,Variable + \mu_i + \lambda_t + \varepsilon_{i,t}}$$

where μ_i represents paper fixed effect, λ_t represents year fixed effect, and $\varepsilon_{i,t}$ is the error term.

We developed two models for our analysis. Model 1, the control model, incorporates only the control variables, while Model 2, our full model, includes both control and main independent variables. The performance of these models is detailed in Table 2, where we present the estimated coefficients and key model fit indices.

To assess model performance, we utilized the log-likelihood and the Akaike Information Criterion (AIC) as comparative metrics (Bozdogan 1987). A higher log-likelihood and a lower AIC are indicative of a better-fitting model. Based on these criteria, Model 2 demonstrates a superior fit compared to the control model, suggesting that the inclusion of main independent variables significantly enhances the model's explanatory power.

According to Model 2 result, Models that occupy central positions and are utilized by other models significantly boost the popularity of associated papers, as evidenced by a positive coefficient ($\beta = 2.84$, $p < 0.001$). This suggests that focal models that are integral to other developments within the field tend to enhance the visibility and citation likelihood of the papers in which they are featured, supporting our first hypothesis. Similarly, the location of most central offspring models also positively impacts paper popularity ($\beta = 1.090$, $p < 0.001$), indicating that even models not officially recognized within the papers can significantly influence their diffusion and recognition. Conversely, when considering the average in-degree eigenvector centrality of all connected offspring models, a negative impact on paper citations is observed ($\beta = -1.151$, $p < 0.001$). This could be due to an over-saturation effect where the prominence and distinctiveness of any single model are diluted amid a crowd of centrally located models, diverting attention from the original paper. Additionally, as more work builds on the initial findings, the field may become saturated, causing newer studies to reference more recent derivatives or even bypass the foundational paper if they are deemed more relevant. This suggests that while standout offspring models significantly boost paper citations, the cumulative average influence of many well-connected models might compete for attention, thereby diminishing the overall boost to citation counts.

Variables	**Model 1**	**Model 2**
Main Variable		
Focal Model Location	-- --	2.840 *** (0.075)
Offspring Model Most Central Location	-- --	1.090 *** (0.020)
Offspring Model Average Location	-- --	-1.151 *** (0.021)
Control Variable		
Focal Model Dependency	0.025 *** (0.006)	0.019 ** (0.006)
Offspring Model Max Dependency	-0.514*** (0.015)	-0.669 *** (0.016)
Offspring Model Average Dependency	0.448 *** (0.014)	0.609 *** (0.015)
Last Year Citation	0.005 *** (0.001)	0.008 *** (0.001)
Average Number of Authors Published Work	0.064 *** (0.010)	0.105 *** (0.010)
Average Citations of Authors	0.198 *** (0.005)	0.156 *** (0.005)
Max Number of Authors Published Work	-0.105*** (0.012)	-0.128 *** (0.012)
Max Citations of Authors	-0.088 *** (0.008)	-0.083 *** (0.008)
Observations	227,955	227,955
Log-likelihood	-126208.8	-124436.6
AIC	252433	248895.2
*Note: *p<.05; **p<.01; ***p<.001. Papers fixed effect and year fixed effects are included*		

Table 2. Model Result

Robustness Checks

To ensure the robustness of our analysis, we have implemented a stratification based on papers' cumulative citation counts as of the end of 2023. We categorize papers falling within the top 0.1%, 0.5%, and 1% of the cumulative citation distribution as outliers (Kang et al. 2023). The citation counts corresponding to these percentiles are 3,152.876, 998, and 612, respectively. These outliers comprise 65, 320, and 640 papers, corresponding to 459, 2040, and 3920 observations, respectively. By applying these multiple thresholds for excluding outliers, we aim to demonstrate the robustness and consistency of our analysis.

Our robustness checks, presented in Table 3, show a consistent positive influence on the location of focal models, reinforcing our main results. However, the impact of the location of the most central offspring models appears to change slightly when we exclude papers in the top 1% of citations. After removing these top-star papers with numerous citations, the number of offspring models based on the focal models decreases significantly. This shift is attributed to the high correlation—over 0.9—between the average and maximum locations of the offspring models. When we retain only the average location of the offspring models, the coefficient turns negative once again, lending further support to our hypothesis.

	Remove 0.1%		**Remove 0.5%**		**Remove 1%**			
Main variable	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Focal Model Location	5.644 ***	0.097	6.862 ***	0.127	5.973 ***	0.149	6.094***	0.148
Offspring Model Most Central Location	1.237 ***	0.026	0.504 ***	0.036	-0.337 ***	0.045	--	--
Offspring Model Average Location	-1.316 ***	0.026	-0.577 ***	0.037	0.268***	0.046	-0.078***	0.047
Observation	227496		225906		224026		224026	
Log-Likelihood	-112516.400		-88243.680		-74962.550		-74991.12	

Note: *p<.05; **p<.01; ***p<.001. Papers fixed effect and year fixed effects are included, control variables are omitted in the table due to space limit

Table 3. Robustness Check Result

Discussion and Future Direction

Our investigation, using AI research papers as a proxy, has underscored the significance of both the model dependency network and the dichotomy between focal and offspring models in shaping the diffusion of AI innovation. The position of a model—whether focal or offspring—appears to uniquely influence its prominence within the AI community. This finding suggests that a model's impact is not just a product of its isolated achievements but also a reflection of its role within the broader tapestry of AI innovation. The intricate ties between the popularity of academic papers and the attributes of the models they feature indicate that the practical applications, originality, and theoretical advancements inherent in these models are tightly woven into the fabric of AI evolution. Understanding this complex interplay is paramount to grasping how AI innovation propagates throughout scholarly work.

Looking forward, we aim to delve into the dynamics between focal and offspring models. Particularly, we are interested in how the central positioning of offspring models might influence or even augment the role of focal models. This inquiry will extend to examining whether such positioning can catalyze the advancement of AI innovation in new directions. We will also explore the interaction effects between authors and models to obtain a more comprehensive understanding of the social and technical factors driving innovation. Additionally, we plan to investigate the existence of cross-lagged relationships which may illuminate the temporal sequencing of influence between model prominence and paper popularity.

Moreover, we propose to enrich our analytical framework by incorporating measures such as the Hyperlink-Induced Topic Search (HITS) algorithm, which could reveal the 'hubs' and 'authorities' within the model network and provide a more nuanced understanding of influence within the domain (Kleinberg, 1999). We will also improve our GNN model to increase its practical application. By advancing our methodological approach, we hope to uncover deeper layers of interaction that shape the innovation diffusion process in AI research.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. (SES-2120540). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. 2022. "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics* (40:S1), pp. S293-S340.

Agarwal, C., Queen, O., Lakkaraju, H., and Zitnik, M. 2023. "Evaluating Explainability for Graph Neural Networks," *Scientific Data* (10:1), p. 144.

Alekseeva, L., Azar, J., Gine, M., Samila, S., and Taska, B. 2021. "The Demand for Ai Skills in the Labor Market," *Labour economics* (71), p. 102002.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., and Ha, V. 2018. "Construction of the Literature Graph in Semantic Scholar," *arXiv preprint arXiv:1805.02262*).

Arthur, W. B. 2009. *The Nature of Technology: What It Is and How It Evolves*. Simon and Schuster.

Baek, S., Lee, H., and Kim, H. 2020. "Analysis of Artificial Intelligence's Technology Innovation and Diffusion Pattern: Focusing on Uspto Patent Data," *The Journal of the Korea Contents Association* (20:4), pp. 86-98.

Bhattarai, P., Ghassemi, M., and Alhanai, T. 2022. "Open-Source Code Repository Attributes Predict Impact of Computer Science Research," *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries*, pp. 1-7.

Boland Jr, R. J., Lyytinen, K., and Yoo, Y. 2007. "Wakes of Innovation in Project Networks: The Case of Digital 3-D Representations in Architecture, Engineering, and Construction," *Organization science* (18:4), pp. 631-647.

Bonaccorsi, A., and Rossi, C. 2003. "Why Open Source Software Can Succeed," *Research policy* (32:7), pp. 1243-1258.

Bonacich, P. 2007. "Some Unique Properties of Eigenvector Centrality," *Social networks* (29:4), pp. 555-564.

Bonneel, N., Coeurjolly, D., Digne, J., and Mellado, N. 2020. "Code Replicability in Computer Graphics," *ACM Transactions on Graphics (TOG)* (39:4), pp. 93: 91-93: 98.

Bornmann, L., and Daniel, H. D. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behavior," *Journal of documentation*).

Bozdogan, H. 1987. "Model Selection and Akaike's Information Criterion (Aic): The General Theory and Its Analytical Extensions," *Psychometrika* (52:3), pp. 345-370.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. 2020. "Language Models Are Few-Shot Learners," *Advances in neural information processing systems* (33), pp. 1877-1901.

Brynjolfsson, E., and Kemerer, C. F. 1996. "Network Externalities in Microcomputer Software: An Econometric Analysis of the Spreadsheet Market," *Management science* (42:12), pp. 1627-1647.

Cabanes, C., Grouazel, A., von Schuckmann, K., Hamon, M., Turpin, V., Coatanoan, C., Guinehut, S., Boone, C., Ferry, N., and Reverdin, G. 2012. "The Cora Dataset: Validation and Diagnostics of Ocean Temperature and Salinity in Situ Measurements," *Ocean Science Discussions* (9:2), pp. 1273-1312.

Chen, C. 2012. "Predictive Effects of Structural Variation on Citation Counts," *Journal of the American Society for Information Science and Technology* (63:3), pp. 431-449.

Choi, J., and Yoon, J. 2022. "Measuring Knowledge Exploration Distance at the Patent Level: Application of Network Embedding and Citation Analysis," *Journal of informetrics* (16:2), p. 101286.

Chung, L., Nixon, B. A., Yu, E., and Mylopoulos, J. 2012. *Non-Functional Requirements in Software Engineering*. Springer Science & Business Media.

Consul, P., and Famoye, F. 1992. "Generalized Poisson Regression Model," *Communications in Statistics-Theory and Methods* (21:1), pp. 89-109.

Cosentino, V., Izquierdo, J. L. C., and Cabot, J. 2017. "A Systematic Mapping Study of Software Development with Github," *Ieee access* (5), pp. 7173-7192.

Cummings, D., and Nassar, M. 2020. "Structured Citation Trend Prediction Using Graph Neural Networks," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: IEEE, pp. 3897-3901.

Davenport, T. H., and Ronanki, R. 2018. "Artificial Intelligence for the Real World," *Harvard business review* (96:1), pp. 108-116.

Fan, Y., Ju, M., Zhang, C., and Ye, Y. 2022. "Heterogeneous Temporal Graph Neural Network," *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*: SIAM, pp. 657-665.

Fleming, L., Mingo, S., and Chen, D. 2007. "Collaborative Brokerage, Generative Creativity, and Creative Success," *Administrative science quarterly* (52:3), pp. 443-475.

Frank, M. R., Wang, D., Cebrian, M., and Rahwan, I. 2019. "The Evolution of Citation Graphs in Artificial Intelligence Research," *Nature Machine Intelligence* (1:2), pp. 79-85.

Fredström, A., Wincent, J., Sjödin, D., Oghazi, P., and Parida, V. 2021. "Tracking Innovation Diffusion: Ai Analysis of Large-Scale Patent Data Towards an Agenda for Further Research," *Technological Forecasting and Social Change* (165), p. 120524.

Fricke, S. 2018. "Semantic Scholar," *Journal of the Medical Library Association: JMLA* (106:1), p. 145.

Gao, K., Yoo, Y., and Schecter, A. 2021. "A Dynamic Analysis of the Complex Interplay between Internal and External Networks of Open-Source Projects on Innovation Behaviors,").

Giles, C. L., Bollacker, K. D., and Lawrence, S. 1998. "Citeseer: An Automatic Citation Indexing System," *Proceedings of the third ACM conference on Digital libraries*, pp. 89-98.

Holm, A. N., Plank, B., Wright, D., and Augenstein, I. 2020. "Longitudinal Citation Prediction Using Temporal Graph Neural Networks," *arXiv preprint arXiv:2012.05742*).

Hou, M., Ren, J., Zhang, D., Kong, X., Zhang, D., and Xia, F. 2020. "Network Embedding: Taxonomies, Frameworks and Applications," *Computer Science Review* (38), p. 100296.

Kang, D., Kang, T., and Jang, J. 2023. "Papers with Code or without Code? Impact of Github Repository Usability on the Diffusion of Machine Learning Research," *Information Processing & Management* (60:6), p. 103477.

Kannan, A. V., Fradkin, D., Akrotirianakis, I., Kulahcioglu, T., Canedo, A., Roy, A., Yu, S.-Y., Arnav, M., and Al Faruque, M. A. 2020. "Multimodal Knowledge Graph for Deep Learning Papers and Code," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3417-3420.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., and Cohan, A. 2023. "The Semantic Scholar Open Data Platform," *arXiv preprint arXiv:2301.10140*).

LaHuis, D. M., Hartman, M. J., Hakoyama, S., and Clark, P. C. 2014. "Explained Variance Measures for Multilevel Models," *Organizational Research Methods* (17:4), pp. 433-451.

Li, J., Zhang, C., and Zhang, C. 2023. "Heterogeneous Temporal Graph Neural Network Explainer," *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1298-1307.

Li, X., Chen, H., Huang, Z., and Roco, M. C. 2007. "Patent Citation Network in Nanotechnology (1976–2004)," *Journal of Nanoparticle Research* (9), pp. 337-352.

Lin, J., Wang, Y., Yu, Y., Zhou, Y., Chen, Y., and Shi, X. 2022. "Automatic Analysis of Available Source Code of Top Artificial Intelligence Conference Papers," *International Journal of Software Engineering and Knowledge Engineering* (32:07), pp. 947-970.

Makridakis, S. 2017. "The Forthcoming Artificial Intelligence (Ai) Revolution: Its Impact on Society and Firms," *Futures* (90), pp. 46-60.

Mueller-Langer, F., Fecher, B., Harhoff, D., and Wagner, G. G. 2019. "Replication Studies in Economics—How Many and Which Papers Are Chosen for Replication, and Why?," *Research Policy* (48:1), pp. 62-83.

Nanumyan, V., Gote, C., and Schweitzer, F. 2020. "Multilayer Network Approach to Modeling Authorship Influence on Citation Dynamics in Physics Journals," *Physical Review E* (102:3), p. 032303.

Neville, J., and Jensen, D. 2007. "Relational Dependency Networks," *Journal of Machine Learning Research* (8:3).

Petersen, A. M., Jung, W.-S., Yang, J.-S., and Stanley, H. E. 2011. "Quantitative and Empirical Demonstration of the Matthew Effect in a Study of Career Longevity," *Proceedings of the National Academy of Sciences* (108:1), pp. 18-23.

Pósfai, M., and Barabási, A.-L. 2016. *Network Science*. Citeseer.

Price, D. d. S. 1976. "A General Theory of Bibliometric and Other Cumulative Advantage Processes," *Journal of the American society for Information science* (27:5), pp. 292-306.

Priem, J., Piwowar, H., and Orr, R. 2022. "Openalex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," *arXiv preprint arXiv:2205.01833*).

Ren, X., Liu, J., Yu, X., Khandelwal, U., Gu, Q., Wang, L., and Han, J. 2014. "Cluscite: Effective Citation Recommendation by Information Network-Based Clustering," *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 821-830.

Rogers, E. M. 2010. *Diffusion of Innovations*. Simon and Schuster.

Saputro, D. R. S., Susanti, A., and Pratiwi, N. B. I. 2021. "The Handling of Overdispersion on Poisson Regression Model with the Generalized Poisson Regression Model," *AIP Conference Proceedings*: AIP Publishing.

Schares, E., and Mierz, S. 2023. "Using Openalex to Analyse Cited Reference Patterns," *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*: International Conference on Science, Technology and Innovation Indicators.

Scheidsteger, T., and Haunschild, R. 2023. "Which of the Metadata with Relevance for Bibliometrics Are the Same and Which Are Different When Switching from Microsoft Academic Graph to Openalex?," *Profesional de la información/Information Professional* (32:2).

Shao, H., Sun, D., Wu, J., Zhang, Z., Zhang, A., Yao, S., Liu, S., Wang, T., Zhang, C., and Abdelzaher, T. 2020. "Paper2repo: Github Repository Recommendation for Academic Papers," *Proceedings of The Web Conference 2020*, pp. 629-639.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search," *nature* (529:7587), pp. 484-489.

Stojnic, R., Taylor, R., Kardas, M., Kerkez, V., and Viaud, L. 2022. "Papers with Code-the Latest in Machine Learning," *URL: https://paperswithcode. com*).

Takeda, Y., and Kajikawa, Y. 2010. "Tracking Modularity in Citation Networks," *Scientometrics* (83:3), pp. 783-792.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. 2015. "Line: Large-Scale Information Network Embedding," *Proceedings of the 24th international conference on world wide web*, pp. 1067-1077.

Tang, X., Li, X., Ding, Y., Song, M., and Bu, Y. 2020. "The Pace of Artificial Intelligence Innovations: Speed, Talent, and Trial-and-Error," *Journal of Informetrics* (14:4), p. 101094.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. 2013. "Atypical Combinations and Scientific Impact," *Science* (342:6157), pp. 468-472.

Uzzi, B., and Spiro, J. 2005. "Collaboration and Creativity: The Small World Problem," *American journal of sociology* (111:2), pp. 447-504.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," *Advances in neural information processing systems* (30).

Wang, Y., Shi, Z., Richardson, T., Huang, K., Weerawarna, P., and Wang, Y. 2023. "Building Explainable Graph Neural Network by Sparse Learning for the Drug-Protein Binding Prediction," *bioRxiv*), p. 2023.2008. 2028.555203.

White, H. D. 2001. "Authors as Citers over Time," *Journal of the American Society for Information Science and Technology* (52:2), pp. 87-108.

Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., and Hsieh, C.-Y. 2023. "Chemistry-Intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking," *Nature Communications* (14:1), p. 2585.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. 2018. "How Powerful Are Graph Neural Networks?," *arXiv preprint arXiv:1810.00826*).

Yan, E., and Ding, Y. 2009. "Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis," *Journal of the American Society for Information Science and Technology* (60:10), pp. 2107-2118.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. 2019a. "Gnn Explainer: A Tool for Post-Hoc Explanation of Graph Neural Networks," *arXiv preprint arXiv:1903.03894* (8).

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. 2019b. "Gnnexplainer: Generating Explanations for Graph Neural Networks," *Advances in neural information processing systems* (32).

Zhang, H., Wu, B., Yuan, X., Pan, S., Tong, H., and Pei, J. 2024a. "Trustworthy Graph Neural Networks: Aspects, Methods, and Trends," *Proceedings of the IEEE*).

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., and Huang, Y. 2024b. "Missing Institutions in Openalex: Possible Reasons, Implications, and Solutions," *Scientometrics*), pp. 1-23.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. 2020. "Graph Neural Networks: A Review of Methods and Applications," *AI open* (1), pp. 57-81.

Zingg, C., Nanumyan, V., and Schweitzer, F. 2020. "Citations Driven by Social Connections? A Multi-Layer Representation of Coauthorship Networks," *Quantitative Science Studies* (1:4), pp. 1493-1509.