*[E*ffect of Science-Industry Interaction and Collaborative Structures on the Research Performance of Academic Scientists. The Case of Artificial Intelligence in Canada*]*

[Niushin Khamsehli]

A thesis in the
Department of Quality System Engineering

Presented in Partial Fulfillment of the Requirements for the Degree of
Master of Applied Science (MASc), at
Concordia University

Montréal, Québec, Canada

[April, 2026]

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **[Niushin Khamsehli]**

Entitled: **[Effect of Science-Industry Interaction and Collaborative Structures on the Research Performance of Academic Scientists. The Case of Artificial Intelligence in Canada]**

and submitted in partial fulfillment of the requirements for the degree of

[Master of Applied Science (MASc)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

________________________________ Chair
[Dr. Anjali Awasthi]

________________________________ Examiner
[Dr. Arash Mohammadi]

________________________________ Supervisor
[Dr. Andrea Schiffauerova]

Approved by ____

[Chair of Department or Graduate Program Director]

[04/22/2026] ____

[Dean of Faculty]

ABSTRACT

[Effect of Science-Industry Interaction and Collaborative Structures on the Research Performance of Academic Scientists.

The Case of Artificial Intelligence in Canada]

[Niushin Khamsehli]

Science-Industry interaction accelerates scientific advancement by providing researchers with access to practical challenges, resources and applied perspectives. At the same time, industry benefits from cutting-edge academic knowledge that supports technical innovation. Because of these mutual advantages, such interactions have received increasing attention from researchers in recent years. The overarching goal of this thesis is to identify the factors that significantly influence this interaction. To measure the extent of science-industry interaction, this study relies on two methods: co-authorship between academic and industrial scientists and author-inventors. We analyze how several behavioral and structural factors-including group size, repeated collaboration, industrial participants, dual-status scientists and network characteristics such as betweenness, degree, and average coauthor cohesion influence the research performance of academic scientists in terms of productivity and citation impact. This analysis is based on publication data from Scopus and patent from Lens and employs a Poisson regression model to analyze publication counts and an Ordinary Least Square (OLS) regression model to examine average citation impact.

Despite several methodological limitations—such as challenges in author name disambiguation, the use of cross-sectional data, and the focus on a specific disciplinary domain—this study offers important empirical insights into how collaboration strategies and network structures shape both the productivity and impact of scientific research. The results show that brokerage positions in collaboration networks, measured by betweenness centrality, positively affect scientific productivity, although the relationship follows an inverted-U pattern, suggesting diminishing benefits beyond a certain level. Similarly, degree centrality positively influences citation impact but also exhibits a nonlinear pattern. The findings further highlight the importance of industry engagement: a higher percentage of industry collaborators is associated with increased publication output, while the presence of industry collaboration is linked to higher citation impact. In addition, author–inventor status shows a positive relationship with both productivity and citation outcomes, indicating the benefits of combining scientific and technological activities. Importantly, the results reveal a significant interaction between industry collaboration and collaboration networks, suggesting that the citation benefits of industry engagement become stronger for researchers with broader collaboration networks. Overall, the findings demonstrate that the interaction between academic collaboration structures and industry engagement plays a crucial role in shaping scientific performance.

Acknowledgements

I would like to extend my heartfelt appreciation to all those who supported me throughout the course of this thesis. I am especially grateful to my supervisor, Dr. Schiffauerova, for her invaluable guidance, insightful feedback, and unwavering support. Her mentorship played a crucial role in shaping the direction and quality of this work.

I am deeply thankful to my parents for their unconditional love, encouragement, and patience during every step of this journey. Their constant belief in me has been a source of strength throughout my academic path.

A very special thanks goes to my son, Artin, whose presence fills my life with joy and purpose. His smile has been my greatest motivation through challenging times, and I dedicate this achievement to him with all my love.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction and Motivation

Science–industry collaboration plays a crucial role in shaping modern scientific research (Bianchini et al., 2020), particularly in fast-evolving fields such as artificial intelligence. Such collaboration provides academic scientists with access to practical challenges, complementary expertise, and advanced technological resources that are often unavailable within university settings (Santoro et al., 2019). These interactions enrich the research process by fostering new scientific questions, improving data quality, and promoting knowledge flows that enhance both the productivity and the visibility of academic work (Fortunato et al., 2018). As a result, the significance of this collaboration lies not only in its broader contribution to innovation systems but also in its direct influence on the quality, relevance, and impact of scientific outputs.

A wide range of methodological approaches has been used to examine collaboration between science and industry, reflecting the diversity of interaction channels and the complexity of knowledge exchange. Some studies rely on co-authorship analysis to capture formal collaboration patterns (Katz & Martin, 1997), while others use patent data to identify academic inventors involved in industrial innovation (Meyer, 2003). Additional approaches analyze contract research and industry-funded projects (Bekkers & Bodas Freitas, 2008) or collect survey-based evidence to capture informal knowledge-exchange practices (Perkmann et al., 2013). Similarly, studies evaluating the performance of academic scientists employ various bibliometric indicators, such as publication counts (Stephan, 1996), citation-based metrics (Van Raan, 2005), the h-index (Hirsch, 2005), and journal quality (Lee & Bozeman, 2005). Together, these studies demonstrate the multidimensional nature of academic performance and the multiple pathways through which collaboration can shape scientific outcomes.

Despite extensive research on science–industry collaboration, prior studies often treat industry engagement as having a uniform and additive effect on academic performance (Abramo et al. 2019). Most analyses focus on single forms of interaction—such as co-authorship (Hottenrott & Lawson, 2017) or patenting (Abdullah & Hicks, 2019)—and evaluate their impact using one primary performance indicator (Hicks et al., 2015).

This fragmented perspective overlooks the possibility that the effects of industry collaboration may vary depending on the structural and relational contexts in which scientists operate. In particular, the impact of industry engagement may depend on scientists' positions within collaboration networks and the structural characteristics of their research teams. Well-connected researchers or those occupying brokerage positions may be better positioned to leverage industry ties into high-impact publications, while others may obtain more limited returns. However, this possibility remains underexplored and warrants systematic empirical investigation Similarly, team-level features—such as size, disciplinary diversity, repeated collaboration, and team embeddedness—are likely to shape how effectively industry collaboration translates into scientific productivity and citation impact, yet this moderating role has received limited systematic attention

Moreover, much of the existing literature assumes a linear relationship between collaboration intensity and research performance (Leahey et al., 2017). Such an assumption neglects the possibility of nonlinear dynamics, including diminishing returns, threshold effects, or inflection points in the relationship between industry engagement and academic outcomes.

Addressing these gaps, this study develops an author-level empirical framework that models interaction effects between industry engagement (measured through the proportion or presence of industry involvement as well as status of author-inventor) and key structural characteristics, including network embeddedness and team-level features. By allowing the effect of industry collaboration to vary across structural contexts and testing for potential nonlinear relationships, the study provides a conditional and multidimensional assessment of how science–industry interaction shapes publication quantity and citation impact.

This study focuses on the artificial intelligence research ecosystem in Canada. In this analysis, data on scientific publications, and inventor records are integrated to examine the effects of collaboration between academic and industrial researchers. Given that artificial intelligence has become one of the most influential and widely applied technological fields in recent decades, understanding the dynamics of collaboration in this domain is particularly important. The findings of this study can contribute to forming a better interaction between academic scientists and industry researchers and may provide useful insights for strengthening science–industry collaboration within the Canadian research ecosystem.

Chapter 2: Literature Review

2.1 Introduction

In this chapter, the literature examining the effects of academic–industry collaboration, collaboration network structures, and the behavioral and structural characteristics of scientists on scientific performance is reviewed. This chapter covers three main areas: methods for measuring academic–industry interaction, structural network measures and other collaboration-related independent variables (such as industry involvement and repeated collaboration patterns), and approaches used to evaluate scientific performance. The chapter concludes by identifying the research gaps in the existing literature and outlining the objectives of this thesis.

2.2 Methods of Measurement of Academic–Industry Interaction

In recent decades, a growing body of literature has emerged around the question of how to evaluate and quantify interactions between academia and industry. As these collaborations increasingly influence research agendas, funding priorities, and national innovation policies, scholars have emphasized the importance of developing reliable methods for measuring science–industry linkages (Perkmann et al., 2013; Abramo et al., 2011). Accurately identifying these interactions is crucial not only for understanding the structure of research networks, but also for assessing their impact on knowledge transfer, technological advancement, and scientific performance (Meyer, 2006). Researchers have adopted a variety of approaches—ranging from author-level matching to citation analysis and affiliation tracking—each offering distinct advantages and limitations depending on the context and available data (Perkmann et al, 2013).

2.2.1 Author–Inventor Method

The author–inventor method seeks to identify individuals who are simultaneously active in both academic publishing and patenting. This approach matches names across scientific publication databases and patent records to detect dual roles and is widely used as a proxy for direct engagement in both knowledge production and technological application. Balconi et al.

(2004) were among the first to apply this method by linking data from the European Patent Office and bibliometric sources. Similarly, Czarnitzki et al. (2007) used this approach to study academic entrepreneurship in Germany.

One major advantage of the author–inventor method is its ability to reveal researchers who bridge the gap between academia and industry (Azoulay et al., 2019)—those who not only generate scientific knowledge but also contribute directly to technological innovation. This offers a person-level measure of science–industry linkage.

However, the method suffers from the challenge of name disambiguation (Torvik & Smalheiser, 2009), particularly in the absence of unique researcher identifiers such as ORCID, common names, inconsistent use of initials, and changes in affiliation may lead to false matches or missed connections (Carayol & Cassi, 2009). In some cases, researchers may also use different naming conventions in academic versus patenting contexts, further complicating the matching process (Bianchini et al., 2020).

2.2.2 Academic–Industrial Co-publication

Another widely used method involves identifying scientific publications co-authored by individuals affiliated with both academic and industrial institutions. The typical process involves extracting affiliation data from each publication and classifying each co-author's institutional affiliation as either "academic" or "industrial." Co-publications are then identified when at least one academic, and one industrial affiliation appear on the same paper. This method assumes that co-authorship across academic and industrial affiliations serves as a proxy for formal inter-sectoral collaboration (Perkmann et al., 2013; Civera et al., 2020). The co-authorship approach has been widely adopted in studies examining university–industry collaboration patterns. For example, Perkmann et al. (2013) used bibliometric co-publication data to assess the extent and determinants of academic–industry research collaboration. More recently, Abramo et al. (2019) and Civera et al. (2020) employed affiliation-based co-authorship measures to analyze institutional and national-level collaboration dynamics. This approach is particularly useful for analyzing collaboration patterns at the organizational and institutional levels, where affiliation-based data allow for systematic classification of inter-sectoral linkages (Perkmann et al., 2013; Abramo et al., 2019).

A key advantage of this method is its data availability and simplicity—affiliation metadata is often readily available in bibliometric databases, and classification can be automated at scale (Waltman, 2016). Furthermore, this method captures active, intentional collaboration, especially in fields where joint funding or applied problem-solving drives co-authorship (Perkmann & Walsh, 2007).

However, several challenges affect its precision. First, some institutions function in both academic and industrial roles (e.g., public–private research centers), making classification ambiguous (Ranga & Etzkowitz, 2013). Second, not all industrial partners list corporate affiliations explicitly, particularly when involved informally or through consulting roles (Bianchini et al., 2020). Third, organizational names may appear inconsistently, especially in multi-branch or multinational firms, leading to under- or over-estimation of collaboration rates (D'Angelo et al., 2011). Despite these limitations, academic–industrial co-publication remains one of the most direct and scalable indicators of institutional collaboration and is often used in policy analysis and cross-country benchmarking (Perkmann et al., 2013).

2.2.3 Citation-Based Analysis

This method tracks citations from patents to scientific literature as a proxy for knowledge flow from academia to industry. Citations embedded in patent documents are analyzed to identify references to prior academic publications, which suggests that the patented technology may have been influenced by scientific research. Several recent studies have applied patent-to-publication citation analysis to examine knowledge flows between academia and industry. For example, Ahmadpoor and Jones (2017) analyzed large-scale U.S. patent citations to scientific publications to map the distance between science and technology. More recently, Marx and Fuegi (2020) used linked patent–publication datasets to trace scientific influence on industrial innovation at the global level.

The main strength of citation-based analysis is that it reveals indirect connections, capturing influence even in the absence of formal collaboration (Ahmadpoor & Jones, 2017). This method can be especially valuable in identifying the scientific foundations of industrial innovation (Fleming & Sorenson, 2004).

However, citation formatting inconsistencies in patents pose a significant challenge. References may be incomplete or presented in formats that complicate accurate linking to academic databases (Jefferson et al., 2018). Additionally, the presence of a citation does not necessarily imply meaningful influence or collaboration, as citations may reflect examiner additions or general background references (Criscuolo & Verspagen, 2008).

2.2.4 University-Based Patenting

The university-based patenting approach involves analyzing patents that are either filed by universities or that list university-affiliated inventors. It is often used to assess the technological output and commercialization potential of academic research. Azagra-Caro et al. (2006) applied this method to Spanish universities, while Geuna and Nesta (2006) used it for cross-country comparisons.

The key advantage of this approach is that it provides clear evidence of academic contributions to innovation, particularly when universities maintain active technology transfer offices (Secundo et al., 2017). It is also helpful for tracking institutional engagement in patenting over time (Guerrero et al., 2015).

Nonetheless, this method underrepresents informal or third-party commercialization, as some university researchers may patent independently or through spin-offs (Fini et al., 2018). Variability in how university affiliations are recorded in patent documents also limits the accuracy of institutional mapping (Magerman et al., 2015).

2.2.5 Scientific Publications by Industrial R&D Staff

The scientific publications by industrial R&D staff method focuses on identifying scientific articles authored by researchers working in corporate R&D divisions. It helps assess the scientific output of private-sector innovation ecosystems. Godin (1996) used this method to analyze Canadian industrial R&D publications, and Belkhodja and Landry (2007) extended the approach in studies of public–private knowledge exchange.

The benefit of this approach is that it allows researchers to quantify the scientific engagement of industry, an often-overlooked contributor to research production (Bikard et al., 2019). It provides insight into how deeply industry participates in academic discourse.

However, identifying industrial affiliations in bibliometric data remains challenging due to inconsistent naming and decentralized corporate structures (Donner et al., 2020). This method also does not directly capture collaboration, but rather reflects output by industry-based researchers, whether collaborative or not (Zhang et al., 2021).

2.3 Structural and Relational Dimensions of Scientific Collaboration

Beyond the mere existence of academic–industry collaboration, the structural and relational configuration of scientific collaboration plays a critical role in shaping research performance. Social network analysis has demonstrated that the position of researchers within collaboration networks influences access to information, resources, and recognition (Freeman, 1979; Wasserman & Faust, 1994). In co-authorship networks, collaboration can be analyzed through researchers' structural positions as well as through relational attributes such as the intensity, diversity, and stability of collaborative ties (Newman, 2001). Network-based indicators—including centrality measures such as degree, betweenness, closeness, and eigenvector centrality—capture the extent to which a scientist is embedded within the collaboration structure and connected to influential actors (Bonacich, 1987). Empirical studies have shown that these structural positions are associated with variations in scientific productivity and citation impact (Lee & Bozeman, 2005; Abbasi et al., 2012). In addition, collaboration-related variables such as industry involvement and repeated partnerships reflect the composition and persistence of collaborative relationships, which may further shape research outcomes. Together, these structural and relational dimensions provide a coherent framework for examining how patterns of collaboration influence scientific performance.

2.3.1 Network

In everyday life, individuals are embedded in various types of networks—those formed through friendships, professional relationships, and family ties. The concept of networks has long played a significant role in engineering, particularly for managing complex systems such as transportation and communication infrastructures. Beginning in the 1960s and 1970s, sociologists also adopted the network framework to analyze social norms, exchanges, and power structures (Cook, 2014). By the 1980s, the network paradigm had emerged as a central analytical tool across the social sciences (Wasserman & Faust, 1994).

2.3.1.1 Network of Co-authorship

A growing number of studies have examined how co-authorship networks facilitate knowledge exchange. These networks are often visualized as graphs where nodes represent scientists and edges denote co-authored publications. Through these collaborative ties, scientists gain more efficient access to knowledge, infrastructure, and funding, which in turn accelerates the production of high-quality outputs such as journal articles, conference papers, and books (Newman, 2001; Ioanid et al., 2018; Guan & Zhao, 2013).

The outcome of such cooperation is typically a jointly authored publication, which serves as a traceable record of shared intellectual effort. As such, co-authorship data extracted from journal databases have become a valuable source for constructing and analyzing scientific collaboration networks (Glänzel & Schubert, 2004; Eslami et al., 2013). These networks capture the structural organization of research communities and enable the identification of central actors and collaborative patterns associated with scientific productivity (Newman, 2001; Fortunato et al., 2018).

Network Centrality Measures

In network analysis, centrality measures are used to quantify the position and relative importance of actors within a network. These measures capture how well-connected an individual is, how strategically positioned they are within the structure, and how much influence they may exert over the flow of information and resources (Freeman, 1979; Borgatti & Everett, 2006). In the context of co-authorship networks, structural properties play a central role in understanding how collaborative patterns influence the scientific performance of researchers (Wang et al., 2021). These properties help identify not only the position and connectivity of individuals within the network but also the dynamics through which information and resources circulate (Chen et al., 2017). Various network metrics have been proposed to quantify these characteristics. Some of these metrics quantify the number of direct and indirect links associated with everyone (node), such as degree centrality and eigenvector centrality (Opsahl et al., 2010). Others, such as betweenness centrality, provide insight into a node's brokerage position, indicating how often a researcher acts as a bridge between otherwise disconnected groups (Freeman, 1979; Abbasi et al., 2012). The following section focuses on the most widely used centrality measures in scientific collaboration networks, particularly within co-authorship networks, and discusses their theoretical foundations and empirical implications.

Degree Centrality

Degree centrality is one of the most basic yet widely used measures in network analysis. It quantifies the number of direct links a node has, and in co-authorship networks, it reflects the number of unique collaborators a researcher is directly connected to. Researchers with high degree centrality are typically more active in collaborative activities and tend to hold more prominent positions within the scientific community.

$$C_D(i) = \sum_{j=1,\, j\neq i}^{n} a_{ij}$$

Where n is the total number of nodes and a_{ij} is the (i, j) entry of the adjacency matrix.

In the co-authorship network, it is equal to 1 if a tie exists between authors i and j, and 0 otherwise (with a_{ii}=0, as self-loops are not allowed). In an undirected co-authorship network, $C_{D(i)}$ corresponds to the number of unique collaborators of author i.

Betweenness Centrality

Betweenness centrality measures the extent to which a node lies on the shortest paths between other pairs of nodes in a network. It quantifies the brokerage potential of a node by capturing how frequently it serves as an intermediary along geodesic paths connecting other nodes (Freeman, 1977; Newman, 2010).

In co-authorship networks, betweenness centrality reflects the extent to which a researcher acts as a bridge between otherwise disconnected groups of collaborators. Researchers with high betweenness centrality often occupy strategic structural positions that facilitate the flow of knowledge across disciplinary or institutional boundaries (Borgatti & Halgin, 2011). Such brokerage positions may provide access to diverse information and enhance visibility, influence, and innovation potential (Abbasi et al., 2011).

Formally, in an undirected network, betweenness centrality of node i is defined as:

$$C_B(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Where:

σ_{st} denotes the total number of shortest paths between nodes s and t,

$\sigma_{st}(i)$ denotes the total number of those shortest paths that pass-through node i.

For comparability across networks of different sizes, the measure is often normalized by dividing by *(n-1) (n-2)/2* in undirected networks.

Several studies have demonstrated that betweenness centrality is positively associated with research visibility, innovation, and scientific impact (Freeman, 1977; Burt, 1992; Abbasi et al., 2012).

Clustering Coefficient

The clustering coefficient measures the extent to which a node's neighbors are connected to each other. It captures the degree of local cohesiveness in a network by assessing how likely it is that two collaborators of a researcher also collaborate with one another (Watts & Strogatz, 1998; Newman, 2010).

In co-authorship networks, a high clustering coefficient indicates that a researcher is embedded in a tightly connected group where collaboration tends to occur within a closed circle. While such cohesive structures can enhance trust, coordination, and efficiency, they may also limit exposure to novel ideas and external collaborators (Uzzi & Spiro, 2005).

Formally, in an undirected and unweighted network, the local clustering coefficient of node *i* is defined as:

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

where:

$\boldsymbol{T_i}$ is the number of triangles involving node *i*,

$\boldsymbol{k_i}$ is the degree of node *i*,

$\boldsymbol{k_i}(\boldsymbol{k_i}$-1)/2 represents the maximum possible number of connections among node *i*'s neighbors. The clustering coefficient ranges between 0 and 1, where higher values indicate stronger local cohesion.

Eigenvector Centrality

Eigenvector centrality measures a node's influence in a network by considering not only the number of its connections but also the importance of the nodes to which it is connected (Bonacich, 1987; Newman, 2010). Unlike degree centrality, which treats all connections equally, eigenvector centrality assigns greater weight to connections with highly central nodes.

In co-authorship networks, this metric captures the extent to which a researcher is connected to other well-connected and influential scholars. A researcher with high eigenvector centrality is embedded in influential parts of the collaboration network and may benefit from enhanced visibility, prestige, and access to high-impact research communities (Borgatti & Everett, 2006).

Formally, eigenvector centrality is defined as:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j$$

or in matrix notation:

$$\mathbf{Ax} = \lambda \mathbf{x}$$

where:

A is the adjacency matrix of the network,

$\boldsymbol{x_i}$is the centrality score of nodes *i*,

λ is the largest eigenvalue of A,

x is the eigenvector corresponding to λ.

The centrality score of a node is therefore proportional to the sum of the centralities of its neighbors.

Effect of Collaboration Network on Scientific Performance

Numerous studies have emphasized the importance of network position in determining scientific productivity. Ahuja (2000), using patent data from U.S.-based R&D firms between 1980 and 1990, demonstrated that scientists with higher degree centrality—i.e., more direct collaborators—tended to produce more patents and scientific articles. The study concluded that more connections provide access to diverse knowledge pools, thereby enhancing productivity However, not all results support the assumption that higher degree centrality enhances scientific performance. Liu et al. (2015), in a study of biomedical research publications from 2000 to 2010 found that extremely high levels of degree centrality were associated with a decline in the average number of citations per article. They argued that overly extensive collaboration may dilute individual contributions and reduce the scientific impact per output. These contrasting findings suggest that the effects of degree centrality may vary across contexts and may not follow a strictly linear pattern.

Similarly, Abbasi et al. (2011) analyzed publication data in the field of information science using co-authorship networks constructed over the 2000–2009 period. They found that betweenness centrality was positively associated with higher h-index scores. Scientists who acted as bridges between otherwise disconnected subgroups tended to benefit from their brokerage positions, potentially accessing novel ideas and collaborations.

The influence of eigenvector centrality, which considers not only the number but also the quality of one’s collaborators, has been explored by Huang et al. (2015) in the context of physics research in China from 2005 to 2012. Their results indicated that scientists with high eigenvector centrality achieved more citations, suggesting that being connected to other influential scientists amplifies research impact.

Clustering coefficient, reflecting the degree to which an author`s collaborators are also connected to one another, has also shown nuanced effects. Zhang et al. (2021), in their study on Chinese AI researchers (2008–2018), found that moderate clustering improved productivity and innovation. However, very high clustering led to reduced diversity of ideas and lower average citation impact, indicating that excessive local cohesion may hinder novel contributions.

Some researchers have looked at combinations of centrality measures. For instance, a study by Kim and Diesner (2016) on sociology publication networks (1998–2013) found that eigenvector centrality did not significantly predict h-index or total citation count when controlling for academic rank and publication volume. This result challenges the assumption that being connected to influential others always leads to higher recognition.

In another study, Tang and He (2018) examined over 12,000 publications in environmental science and showed that researchers with high closeness and eigenvector centrality had higher h-index scores. Their analysis, covering the years 2000–2015, confirmed that being both well-connected and positioned near influential collaborators enhances long-term academic influence.

Overall, the literature confirms that the structural features of collaboration networks significantly influence scientific performance. While most studies support a positive association between centrality measures and output or impact, several findings suggest diminishing or even negative returns beyond certain thresholds. Moreover, different metrics appear to affect different dimensions of performance. These nuanced results highlight the importance of analyzing collaboration networks not just as static structures but as dynamic systems with both opportunities and constraints for scientific achievement.

2.3.2 Relational Characteristics of Scientific Collaboration

While the previous subsection examined the structural position of researchers within collaboration networks, network structure alone does not fully capture the complexity of scientific collaboration. Beyond centrality, the nature, intensity, and stability of collaborative relationships also play a significant role in shaping research outcomes (Granovetter, 1973; Uzzi, 1997).

Prior studies have highlighted the importance of relational collaboration patterns, including repeated partnerships, team size, international collaboration, affiliation diversity, and interdisciplinary collaboration. Repeated collaborations can enhance trust and coordination efficiency, potentially improving research productivity (Gulati, 1995; Lee & Bozeman, 2005). Team size has been associated with both increased knowledge diversity and coordination challenges (Wuchty, Jones, & Uzzi, 2007). International collaboration has been shown to expand access to diverse knowledge pools and is often linked to higher citation impact (Glänzel & Schubert, 2001). Similarly, interdisciplinary collaboration can foster innovation by combining heterogeneous knowledge bases (Rafols & Meyer, 2010). In addition, different forms of interaction with industry and technology may influence the direction and application of research activities (Perkmann et al., 2013; Zhang et al., 2021). These relational characteristics reflect the depth, diversity, and persistence of collaborative ties, which can shape knowledge exchange, resource access, and ultimately scientific performance.

The following subsections review the empirical evidence on these relational dimensions and their documented effects on researchers' productivity and citation impact.

2.3.2.1 Repeated Collaborations

Repeated collaboration, defined as ongoing cooperation between the same researchers across multiple projects or publications, has been shown to significantly influence scientific performance (Gulati, 1995; Lee & Bozeman, 2005). Sustained partnerships may foster familiarity, mutual trust, and coordination efficiency, thereby enhancing productivity and research effectiveness. For example, Eslami et al. (2013), in their study of nanotechnology researchers in Canada (2003–2011), reported a positive association between repeated co-authorship and publication output. Similarly, Bozeman and Corley (2004) found that long-term collaboration improves communication and coordination, contributing to greater research effectiveness.

However, the relationship between repeated collaboration and performance may not be strictly linear. Uzzi and Spiro (2005), in their study of team creativity, identified a curvilinear effect: moderate levels of repeated collaboration enhanced performance, while excessive repetition led to diminishing returns, possibly due to reduced exposure to novel ideas and limited knowledge diversity.

These findings suggest that while sustained partnerships can strengthen coordination and shared expertise, over-reliance on the same collaborators may constrain innovation and reduce the diversity of scientific output.

2.3.2.2 Group Size

Group size, typically measured by the number of co-authors per publication, has been widely studied as a key dimension of collaborative behavior that may influence scientific performance. Larger teams can combine diverse expertise, share resources, and address complex research problems more effectively.

Wuchty, et al., (2007), analyzing a large dataset of publications across disciplines, found that larger teams tend to produce more impactful research, particularly in the natural and life sciences. Similarly, Larivière et al. (2014), using bibliometric data from multiple countries, observed a positive association between team size and citation impact, arguing that broader collaborations often generate more visible and robust outputs.

However, research also highlights potential drawbacks of very large teams. Cummings and Kiesler (2005) found that large collaborative groups may face coordination challenges, communication delays, and uneven workload distribution, which can reduce research efficiency. From a theoretical perspective, increasing team size may generate coordination costs that offset the benefits of knowledge diversity (Katz & Martin, 1997).

These mixed findings suggest that while larger teams can enhance performance through resource pooling and expertise diversity, there may be diminishing returns beyond a certain size, as coordination complexity increases.

2.3.2.3 Team Embeddedness

Team embeddedness refers to the extent to which a research team is situated within a broader collaboration network through repeated ties, shared partners, and overlapping social circles. Highly embedded teams are typically characterized by dense relational structures, strong familiarity among members, and high levels of cohesion, which can facilitate coordination, trust, and efficient knowledge exchange (Granovetter, 1985; Coleman, 1988). In scientific collaboration, such embeddedness can reduce collaboration frictions—such as monitoring costs or communication barriers—because members share reputational information and common norms (Uzzi, 1997). As a result, embedded teams may exhibit higher productivity and smoother execution, particularly when tasks require intensive coordination or tacit knowledge sharing (Reagans & Zuckerman, 2001).

However, the literature also emphasizes that the benefits of embeddedness may come with trade-offs. Dense, tightly connected teams can become socially "closed," limiting exposure to novel information and reducing opportunities to recombine ideas across distant knowledge domains (Burt, 2004; Fleming, Mingo, & Chen, 2007). This aligns with evidence that innovation and high-impact outcomes often depend on a balance between cohesion and access to diverse, non-redundant knowledge sources. In this view, teams that combine internal embeddedness with external bridging ties may be better positioned to generate impactful research by integrating trust-based collaboration with broader informational reach (Uzzi & Spiro, 2005; Phelps, Heidl, & Wadhwa, 2012). Overall, prior studies suggest that team embeddedness can enhance performance through coordination and trust, but excessive embeddedness may lead to redundancy and diminished creative potential—implying a non-linear relationship between embeddedness and scientific performance.

2.3.2.4 Interdisciplinary Collaboration

Interdisciplinary collaboration is defined as scientific cooperation among researchers from different academic fields, and it has been recognized as a significant factor influencing research performance. Research by Leahey et al. (2017), based on U.S. publication data, revealed that interdisciplinary projects often receive more citations, reflecting their broader scientific relevance and appeal across multiple fields. Similarly, Wagner et al. (2011) found that interdisciplinary teams working on global health challenges produced high-impact publications, emphasizing the value of combining diverse knowledge bases to address complex problems. However, some scholars caution that interdisciplinary work can also involve higher cognitive and coordination costs. For example, Rhoten and Parker (2004) argued that differences in terminology, methods, and epistemological assumptions between disciplines may hinder integration and slow down the research process. Despite these challenges, the growing evidence suggests that interdisciplinary collaboration can enhance scientific visibility, foster innovation, and expand the societal relevance of research, particularly when managed effectively.

2.3.2.5 Interaction with Industry

Interaction between academia and industry has gained increasing attention in recent years. Various methods have been employed to identify and measure such collaborations. Among these, the author-inventor approach—which matches individuals who appear both as authors of scientific publications and inventors on patents—has become one of the most widely used indicators of science–industry linkage. Zucker et al. (1998) were among the first to implement this method by linking academic authors in the life sciences with inventors listed in U.S. patent data. They applied a name-matching algorithm supported by affiliation checks and field overlap to disambiguate identities across datasets. Their findings showed that academic scientists who were also inventors had significantly higher publication and citation rates, indicating a strong synergy between engagement in patenting and scientific productivity. Similarly, Azoulay et al. (2019) used full names, affiliations, and temporal co-location to match data from MEDLINE and USPTO. They demonstrated that author-inventors not only published more but were also more cited, attributing the effect to industry exposure and problem-driven research. Balconi et al. (2004) used manual inspection and affiliation cross-referencing to study Italian inventors, finding that author-inventors tended to be more central in academic networks and more prolific in terms of high-impact publications. Lastly, Lissoni et al. (2013) applied multi-step disambiguation on Scopus and EPO data and showed that inventors with academic roles often maintained higher scientific productivity than their non-inventor peers, suggesting that dual engagement enriches knowledge flows and research output.

In parallel, the co-authorship with industry method detects joint publications involving at least one industrial-affiliated author. Schmoch (1997) used institutional affiliation fields from the Science Citation Index to identify such collaborations and found that researchers engaged in industry co-authorship produced more applied and technically cited work. Perianes-Rodriguez and Olmeda-Gomez (2017) implemented a rule-based algorithm using standardized affiliation names and sectoral labels derived from external business directories. Their results showed that academic scientists with frequent industrial co-authorship had higher productivity and broader international visibility. Eslami et al. (2013), focusing on Canadian nanotechnology research, manually labeled affiliations and observed that industry-linked collaborations were associated with increased publication counts and enhanced citation performance. Tijssen (2012) combined Scopus metadata with web-based organizational classification and geographic tagging, concluding that industrial co-authorship contributed to higher impact publications, particularly in engineering and applied sciences.

Another approach to assess science-industry collaboration is the analysis of patent citations to scientific publications. Narin et al. (1997), using data from U.S. patents and the Science Citation Index, examined how frequently patents cited academic research. They found that patents with a higher number of science citations were associated with more innovative and high-value technologies. Their study also showed that scientists whose publications were frequently cited in patents had higher academic impact metrics, such as citations and h-index, suggesting that technological relevance reinforces scholarly visibility. More recently, Meyer (2000) matched patent citations to scientific papers authored by Finnish researchers using a hybrid of manual mapping and bibliometric linkage. He concluded that researchers whose work was cited in patents tended to be more prolific and connected to industry projects, reflecting a mutually reinforcing relationship between academic output and commercial impact.

A further method involves research funding acknowledgments from industry in scientific publications. Crespi and Geuna (2008), analyzing European publications indexed in WoS between 1995 and 2005, examined acknowledgments sections to identify private-sector sponsors. They found that authors who received industry funding had significantly higher productivity and a greater likelihood of publishing in high-impact journals. Their findings suggest that industry-backed research is often oriented toward novel and relevant topics.

Hottenrott and Thorwarth (2011) used natural language processing (NLP) tools to extract sponsor names and categorized them into public and private sources. Based on German data, they found that while both public and industry-funded projects increased output, the citation impact of industry-funded research was notably higher, especially in engineering and applied fields. These studies underscore the importance of financial and institutional links in shaping academic performance through industry engagement.

2.3.2.6 International Collaboration

International collaboration is as co-authorship between researchers affiliated with institutions in different countries, has been widely associated with enhanced scientific performance. Glänzel and Schubert (2001), in a large-scale bibliometric study across multiple disciplines, found that internationally co-authored papers tend to receive significantly more citations compared to domestic collaborations, likely due to their broader dissemination and visibility. Similarly, Wagner and Leydesdorff (2005) emphasized that international networks allow researchers to access complementary expertise and resources, leading to higher-impact and more innovative publications. However, some studies note that the benefits of international collaboration may vary across disciplines and regions. For example, Freeman and Huang (2015), analyzing data from U.S.-based authors, observed that while international collaboration boosts output and citation impact, it can also introduce logistical and communication challenges, especially when time zones, language, and institutional expectations differ. These findings suggest that international partnerships generally contribute positively to performance by expanding intellectual diversity and improving research quality, but they may require more effort to coordinate effectively.

2.4 Measurement and Indicators of Scientific Performance

Evaluation of scientific performance has long been a central concern in research policy and science studies. Traditionally, performance has been assessed using bibliometric indicators derived from publication and citation data, which serve as proxies for productivity, impact, and visibility within the scientific community (Moed, 2005; Waltman, 2016). Among these indicators, publication counts are commonly used to measure research productivity, while citation-based metrics are employed to capture scientific influence and recognition (Bornmann & Daniel, 2008).

Publication output reflects a researcher's capacity to generate and disseminate knowledge. However, simple counts may not account for differences in disciplinary norms, co-authorship practices, or publication types (Larivière et al., 2015). Citation indicators—such as total citations, average citations per paper, and field-normalized citation impact—attempt to address some of these limitations by incorporating measures of scholarly reception and visibility (Waltman et al., 2011). Citations are often interpreted as signals of intellectual influence, although they may also reflect social dynamics, cumulative advantage, and field-specific citation cultures (Merton, 1968; Aksnes et al., 2019).

Beyond basic productivity and citation counts, composite indicators such as the h-index have been proposed to simultaneously capture output and impact (Hirsch, 2005). While widely used, such indices have been criticized for their sensitivity to academic age, disciplinary variation, and skewed citation distributions (Bornmann et al., 2008). Recent research emphasizes the importance of considering the distributional properties of citation data, which are typically highly skewed and overdispersed, making count-based statistical models—such as Poisson or Negative Binomial regression—particularly appropriate for empirical analysis (Cameron & Trivedi, 2013).

In the context of collaboration studies, scientific performance is often operationalized through publication productivity and citation impact at either the paper or author level (Lee & Bozeman, 2005; Wuchty et al., 2007). These measures allow researchers to examine how structural and relational characteristics of collaboration networks shape both the quantity and the quality of research output. Given that collaboration patterns may influence productivity and visibility through different mechanisms, the joint examination of publication counts, and citation-based indicators provides a more comprehensive assessment of scientific performance.

Overall, the literature suggests that no single metric fully captures scientific performance. Instead, a multidimensional approach—incorporating productivity and citation impact—offers a more balanced evaluation framework, particularly when investigating how collaboration behaviors influence research outcomes.

2.5 Research Gap

Despite extensive research on science–industry collaboration, several important gaps remain in understanding how industry engagement shapes academic scientific performance.

First, much of the existing literature treats industry collaboration as having a uniform and additive effect on research outcomes (Abramo et al., 2019). Prior studies often examine whether collaboration with industry increases productivity or citation impact, but they rarely consider how these effects may vary depending on researchers' structural positions within collaboration networks. Although collaboration network research has emphasized structural characteristics such as centrality, brokerage, and embeddedness (Burt, 2004; Uzzi & Spiro, 2005), limited attention has been paid to how these network positions may condition the returns to industry engagement. As a result, it remains unclear whether industry collaboration generates similar productivity and citation outcomes for centrally positioned researchers, highly embedded actors, and more peripheral scientists.

Second, existing research frequently isolates single dimensions of collaboration—such as co-authorship or patenting—without integrating multiple indicators of industry engagement within a unified empirical framework (Hötte & Lawson, 2017; Abdullah & Hicks, 2019). This fragmented approach limits our understanding of how sectoral affiliation, inventor participation, and collaborative network characteristics jointly shape research outcomes.

Moreover, many empirical analyses assume a linear relationship between collaboration intensity and research performance (Leahey et al., 2017). However, theoretical arguments suggest the possibility of diminishing returns, threshold effects, or context-dependent dynamics, particularly when repeated collaboration or strong embeddedness constrains exposure to diverse knowledge sources (Uzzi & Spiro, 2005).

Finally, few author-level studies integrate publication output and citation impact within a single analytical framework in rapidly evolving technological domains such as artificial intelligence. Limited evidence exists on how industry engagement operates within national AI ecosystems, where academic and industrial research are deeply intertwined.

To address these gaps, this study develops an author-level empirical framework that evaluates two complementary dimensions of scientific performance: publication productivity and average citation impact. By modeling publication counts using Poisson regression and estimating citation performance through OLS models based on average citations per author, the study provides a multidimensional assessment of academic outcomes. Furthermore, by incorporating interaction effects between industry engagement and structural network characteristics, the analysis offers a conditional perspective on how science–industry collaboration operates within the Canadian AI research ecosystem.

2.6 Research Objectives

Building on the identified research gaps, this study aims to examine how science–industry interaction operates within the structural and collaborative context of academic researchers in the Canadian artificial intelligence ecosystem. Specifically, the study evaluates scientific performance along two complementary dimensions: publication productivity and average citation impact.

To address the limitations identified in prior research, this thesis develops an author-level empirical framework with the following objectives:

A. To construct multidimensional measures of science–industry interaction by integrating industry co-authorship and authors and inventors overlap at the author level.

B. To examine the direct effect of industry engagement on publication productivity and average citation impact.

C. To analyze how collaborative characteristics—such as repeated collaboration and team size—relate to scientific performance.

D. To evaluate the impact of structural network characteristics—including degree centrality, betweenness centrality, and team embeddedness (mean)—on publication and citation outcomes.

E. To investigate whether the effect of industry engagement on scientific performance varies across structural and collaborative contexts by modeling interaction effects.

F. To assess potential nonlinear relationships between collaboration intensity and research performance.

Chapter 3: Data and Methodology

3.1 Data Collection

This study relies on two primary data sources: one for scientific publications and one for patents, to construct multidimensional measures of science–industry interaction at the author level.

Publication data were retrieved from Scopus, a comprehensive bibliometric database that provides structured metadata on peer-reviewed scientific publications. Scopus offers detailed information on authors' affiliations, publication year, citation counts, and co-authorship relationships. These features are essential for constructing indicators of collaborative structure and scientific performance. Affiliation data enable the identification of whether an author is affiliated with academia, industry, or another sector, which is necessary for measuring industry co-authorship. The availability of citation counts allows for the evaluation of research performance, while structured author metadata supports reliable author-level aggregation.

Patent data were obtained from The Lens database, which integrates global patent records and provides inventor-level information, including full names and, where available, country and institutional affiliation. These attributes are critical for identifying researchers who bridge academic and industrial domains through author–inventor overlap.

The publication dataset covers the period 2015–2020, focusing on artificial intelligence (AI). The patent dataset spans 2000–2024. The broader patent window was selected to account for potential temporal differences between publication and patenting activities, as researchers may publish first and engage in patenting at a later stage.

Both databases were selected due to their structured metadata, international coverage, and suitability for constructing author-level measures of collaboration and science–industry interaction.

3.2 Database Selection

Following the identification of appropriate data sources, a structured selection procedure was applied to construct the analytical sample. The study focuses on Artificial Intelligence (AI) as the field of analysis. AI represents one of the fastest-growing scientific domains, characterized by substantial academic and industrial investment and strong commercialization dynamics. These characteristics make it a suitable empirical context for examining science–industry interaction and collaborative structures.

Relevant publications were retrieved from Scopus using a set of domain-specific keywords, including artificial intelligence, machine learning, deep learning, neural networks, natural language processing, supervised learning, unsupervised learning, and reinforcement learning. These terms were applied to identify publications within the AI domain. The publication dataset covers the period 2015–2020. This time frame was selected to capture a stable phase of expansion in AI research while ensuring temporal consistency and avoiding structural disruptions associated with the COVID-19 period.

To ensure alignment with the research objective, only publications including at least one author affiliated with a Canadian institution were retained. This criterion enables the analysis to focus on researchers operating within the Canadian research system while preserving the international collaborative context of AI research.

Applying the field, temporal, and geographic criteria resulted in a dataset comprising 1,249 unique AI publications published between 2015 and 2020, involving 3,083 unique Canadian authors and 7,951 total authors (including both Canadian and non-Canadian collaborators). These figures represent the structured sample prior to subsequent data cleaning and author-level consolidation procedures.

3.3 Data Cleaning

Following the structured sample construction, several data cleaning and preparation procedures were conducted to ensure accuracy, consistency, and analytical validity at the author level.

Because citation count constitutes one of the two primary indicators of scientific performance in this study, the treatment of missing citation values required careful methodological consideration. Approximately 20% of publication-level citation entries retrieved from Scopus were recorded as missing (NA). Given the central role of citation impact in the empirical analysis, selecting an appropriate treatment strategy was critical to avoid introducing bias or excessive data loss.

Three potential approaches were considered:

A. Removing all authors who had at least one publication with missing citation data.

This approach was rejected because it would have resulted in the exclusion of a substantial proportion of authors from the dataset, leading to significant sample reduction and potential selection bias.

B. Removing only publications with missing citation values while retaining the corresponding authors.

Under this approach, citation-based performance measures would be calculated using only publications with confirmed citation counts.

C. Replacing missing citation values with zero.

This approach assumes that missing entries represent zero citations; however, this assumption cannot be verified and may lead to downward bias in citation-based performance indicators.

Because Scopus records unavailable data as NA, a considerable share of the citation contained missing observations. In this dataset, approximately 20% of citation values were recorded as NA. Assigning arbitrary (zero) value to these missing observations without reliable verification could introduce measurement bias in the regression analysis. Therefore, rather than imputing uncertain value, observations with missing citation information were excluded from the regression dataset.

As a result, the citation-based analytical dataset used for the regression analysis includes 2,448 Canadian authors, compared to 3,083 authors in the initial structured sample. This reduction reflects the removal of observations with unavailable citation information rather than the exclusion of authors from the broader collaboration dataset. In addition to handling missing citation values, author-level consolidation procedures were implemented. Because publications involve multiple co-authors, the dataset was restructured from publication-level records to an author-level analytical format. For each author, publication counts, citation totals, and collaboration-related variables were aggregated across the selected time period (2015–2020). This transformation ensured that each author appears only once in the final analytical dataset.

To support the construction of science–industry interaction measures, affiliation information was carefully standardized to distinguish academic and industry affiliations. This step was necessary for computing industry co-authorship indicators and the percentage of industry collaboration at the author level.

Furthermore, the identification of author–inventor overlap required harmonization of naming conventions across publication and patent datasets. Variations in initials, ordering of names, and formatting inconsistencies were addressed to improve matching accuracy. Matching procedures incorporated name similarity alongside institutional and geographic information to reduce false positives and false negatives in identifying researchers active in both scientific publication and patenting activities. This process resulted in the identification of 150 Canadian author–inventors within the dataset.

Collectively, these data cleaning and restructuring procedures ensured the construction of a consistent, author-level analytical dataset suitable for regression analysis examining the relationship between collaborative structure, science–industry interaction, and scientific performance.

3.4 Methodology

This study adopts a cross-sectional author-level research design, aggregating scientific activity over the period 2015–2020. Rather than modeling year-by-year temporal dynamics, the analysis consolidates publication, citation, and collaboration indicators across a fixed five-year window. This approach allows for the examination of structural and relational patterns of collaboration while maintaining consistency and comparability across authors.

The focus on the pre-pandemic period ensures a stable observational window and avoids structural disruptions associated with the COVID-19 period. Fixed five-year windows are widely used in scientometric research because they are sufficiently long to capture meaningful patterns in publication output and citation accumulation, while remaining short enough to reflect relatively current scientific behavior.

The following subsections describe the construction of the co-authorship network, the operationalization of collaboration and industry interaction variables, and the regression models used to estimate their relationship with scientific performance.

3.4.1 Construction of Scientific Co-authorship Network

The co-authorship network was constructed based on publication data retrieved from Scopus for artificial intelligence (AI) publications published between 2015 and 2020 that included at least one Canadian-affiliated author. The original dataset was structured at the publication level and provided in CSV format, where each record corresponded to a scientific article along with its associated authors and bibliographic information.

To construct an author-level collaboration network, the data were first transformed from publication-level format into an author–publication structure. Duplicate DOI–Author combinations were removed to ensure that each author was uniquely associated with each publication. This step prevented artificial inflation of collaboration ties resulting from repeated records.

Next, all possible pairwise combinations of co-authors within each publication were generated through a self-merge procedure based on the DOI identifier. Self-loops—cases in which an author would be paired with themselves—were explicitly excluded. This process yielded a complete edge list representing co-authorship relations.

The resulting network was modeled as an undirected graph in which each node represents a unique author and each edge represents the existence of at least one joint publication between two authors within the observation window. The graph was constructed programmatically in Python using NetworkX. To ensure accurate author identification and avoid ambiguity due to name duplication, the Scopus Author ID was used as the unique node identifier.

The final network comprises 7,951 unique authors—including both Canadian researchers and their international collaborators—linked through 1,249 AI publications during the 2015–2020 period. Although subsequent statistical analyses focus on Canadian authors, the full co-authorship structure was preserved to maintain the integrity of network-based structural measures.

3.4.2 Operationalization of Collaborative Behavior Variables

Scientific performance is evaluated using two dependent variables: publication count and average citation impact. Because publication count is a non-negative integer variable, Poisson regression models are employed to estimate the effects of collaborative structure and industry interaction on research productivity. Citation impact is measured as the average number of citations per author and treated as a continuous variable; therefore, Ordinary Least Squares (OLS) regression is used for citation-based analyses.

The empirical models examine how structural network position, team-level embeddedness, repeated collaboration, team size, and science–industry interaction is associated with scientific performance. Robust standard errors are applied to account for potential heteroskedasticity.

All models are estimated at the author level, consistent with the cross-sectional research design.

3.4.2.1 Science-Industry Interaction

Science-Industry interaction is captured using two complementary indicators:

Industry Co-authorship

To capture science–industry interaction at the collaboration level, authors were classified into three affiliation-based categories—academic, industrial, and other—based on the institutional affiliation field reported in the Scopus publication records.

The classification procedure relied on keyword-based parsing of affiliation strings. Affiliations containing terms such as “university”, “institute,” or “research center” were categorized as academic. Those including company names or corporate identifiers such as “Inc.”, “Ltd.” or “Corporation” were classified as industrial. Affiliations that did not clearly fall into either category—such as hospitals, government agencies, or non-governmental organizations—were grouped under “other”. Given the focus of this study on science–industry collaboration, authors with dual affiliations (i.e., affiliated with both academic and industrial institutions) were categorized under the industrial group to ensure consistent identification of industry-linked collaboration.

Using the constructed co-authorship network, industry interaction was operationalized at the author level rather than at the publication level. For each focal author, all unique collaborators within the 2015–2020 observation window were identified. The proportion of industrial collaborators was then computed as the number of unique collaborators affiliated with industry divided by the total number of unique collaborators. This measure reflects the structural composition of an author's collaboration network and captures the extent to which a researcher's collaborative ties are embedded in industry-related partnerships.

Author-Inventor Overlap

To identify scientists who are active in both academic publishing and patenting activities, a structured name-matching procedure was conducted between the Scopus (publication) and Lens (patent) databases. The objective was to detect overlapping individuals appearing in both datasets during the study period.

The matching process consisted of four main stages: standardization, name component extraction, candidate selection, and disambiguation.

First, names in both datasets were standardized by converting all characters to lowercase and removing extra spaces, punctuation marks, and special characters to ensure consistency. Given structural differences between the databases—particularly the use of initials for first and middle names in Scopus—special attention was paid to harmonizing name formats across sources.

Second, name components (first name, middle name, and last name) were extracted from both databases. Because Scopus frequently records first and middle names as initials, while Lens often provides full names, a name-expansion strategy was implemented, in the candidate selection phase, exact matching of last names was applied as the primary filter. Subsequently, the first letters of the first or middle name in Lens were required to match the corresponding initial in Scopus. Finally, a name was considered a valid match when the first letters of either the first name or the middle name in Lens corresponded to at least one initial between first name and middle name recorded in the Scopus, regardless of their orders. To reduce false non-matches, the order of first and middle names was treated as non-essential during comparison.

This flexible matching strategy allowed accommodation of variations in name construction and significantly improved matching accuracy.

To enhance precision, full author names were retrieved from Scopus bibliographic records using DOI links when necessary. In addition, obvious typographical variations were considered when identifying equivalent names. For example, minor spelling differences such as “Mohamad” versus “Mohammad”, or “Jon Smith” versus “John Smith”, were treated as referring to the same individual when other identifying information was consistent.

In cases where name correspondences were not immediately clear, additional checks were performed to assess the similarity between name variants and determine whether they referred to the same author. However, due to limited auxiliary identifiers in the Lens database (e.g., incomplete affiliation or geographic information, further verification to identify potential false positive matches was not possible.

Following this procedure, 150 scientists were identified as author–inventors represented in both the Scopus and Lens datasets.

These measures capture distinct dimensions of engagement with industry: collaborative interaction through co-authorship and dual participation in scientific and technological activities.

3.4.2.2 Repeated Collaboration

Repeated collaboration (perc_repeat) captures the extent to which a researcher relies on sustained and stable co-authorship relationships during the 2015–2020 period. Conceptually, it reflects the degree of relational continuity in an author’s collaboration network.

At the author level, this measure is constructed as the proportion of collaboration intensity associated with repeated co-authorship ties. A collaboration is considered repeated only when the same pair of authors co-publishes at least twice within the study window; single co-authorship occurrences are not classified as repeated. The variable is computed by dividing the total frequency of repeated collaboration events (i.e., ties with frequency ≥ 2) by the total frequency of all collaboration events for each researcher. Authors with no collaborative publications are assigned a value of zero.

This variable therefore represents the weighted share of an author’s collaborative activity that occurs within ongoing partnerships rather than one-time interactions.

3.4.2.3 Average Group Size

Average group size (av_gr_size) is constructed at the author level and captures the mean size of collaborative teams in which each researcher participated during the 2015–2020 period. For every publication (DOI), the total number of unique authors is first calculated, including the focal author. The author-level measure is then obtained by summing the group sizes of all publications attributed to a given researcher and dividing this total by the number of publications produced by that researcher within the study window.

This variable therefore reflects the average scale of collaboration characterizing an author's research activity. Because it represents a mean value across multiple publications, it is not restricted to integer values at the author level.

3.4.2.4 Centrality Measures in the Scientific Co-authorship Network

Degree Centrality

Degree centrality is measured within the undirected co-authorship network constructed for the full 2015–2020 period. In this network, nodes represent authors and edges represent the existence of at least one co-authored publication between two researchers during the study window.

Degree centrality is operationalized as the total number of unique co-authors directly connected to a given author in the network. The measure is computed as the raw degree, rather than a normalized centrality score. Consequently, higher values indicate a greater number of distinct collaborative partners.

Because the network is undirected, each collaborative tie is treated symmetrically. This variable therefore captures the structural breadth of a researcher's collaboration network, reflecting the extent of direct connectivity within the scientific community.

Betweenness Centrality

Betweenness centrality is calculated within the undirected co-authorship network constructed for the full 2015–2020 period. In this network, nodes represent authors and edges represent the presence of at least one co-authored publication between two researchers.

Betweenness centrality is computed using the standard shortest-path formulation and is normalized to account for network size. The measure captures the extent to which a given author lies on the shortest paths connecting other pairs of researchers in the collaboration network. Higher values indicate a stronger brokerage position, meaning the author plays a more central role in connecting otherwise distant parts of the network.

The measure is calculated on an unweighted network; thus, it reflects structural positioning rather than collaboration intensity. Authors with no intermediary role in shortest paths receive a value of zero.

The resulting centrality scores were extracted at the author level and merged with the main analytical dataset using the unique author ID identifier. Authors who had no collaborative ties in the network (e.g., solo authors) did not receive centrality scores during computation; therefore, their values were set to zero to retain the full sample of scientists.

3.4.2.5 Team Embeddedness

Empirically, the measure is constructed by first identifying, for each focal author and each publication, the set of co-authors in the publication team. All possible unordered pairs among these collaborators (excluding the focal author) are generated. A pair of collaborators is classified as a repeated pair if the two co-authors have collaborated on more than one publication during the 2015–2020 study period.

For each author–paper observation, the team embeddedness score is computed as the proportion of repeated co-author ties among all possible ties within the collaboration team. The number of possible co-author pairs is defined as

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

where k represents the number of collaborators of the focal author in that publication.

The resulting score ranges from 0 to 1, where higher values indicate a higher level of interconnectedness among the author’s collaborators. Finally, the author-level team embeddedness is obtained by averaging the per-publication embeddedness scores across all publications associated with the focal author during the study period.

3.4.3 Model Specification and Estimation Strategy

Scientific performance is evaluated using two dependent variables: publication count and average citation impact.

Because publication count is a non-negative integer variable, Poisson regression models are employed to estimate the effects of collaborative structure and industry interaction on research productivity. Citation impact is measured as the average number of citations per author and treated as a continuous variable; therefore, Ordinary Least Squares (OLS) regression is used for citation-based analyses.

The empirical models examine how structural network position, team-level embeddedness, repeated collaboration, team size, and science–industry interaction are associated with scientific performance. Robust standard errors are applied to account for potential heteroskedasticity.

All models are estimated at the author level, consistent with the cross-sectional research design.

3.5 Descriptive Visualization of Collaboration Patterns

Before proceeding to the regression analyses, a series of descriptive visualizations were conducted to provide an overview of collaboration patterns among Canadian AI scientists during the 2015–2020 period. These visualizations serve two main purposes. First, they offer a descriptive understanding of the structure and evolution of collaborative activity, including temporal trends, international partnerships, and the distribution of industry involvement. Second, they allow for an initial exploration of potential relationships between collaboration characteristics and research performance indicators.

The figures presented in this section illustrate patterns in publication output, the share of academic–industry collaboration, international co-authorship ties, team size variations across paper types, and differences across productivity and citation groups. Together, these descriptive insights provide contextual grounding for the subsequent econometric analysis and help identify observable trends that motivate the formal modeling strategy.

Figure 1 illustrates the evolution of publication output and the share of industrial collaboration among Canadian AI scientists between 2015 and 2020. The total number of publications shows a clear upward trend over the study period, with a steady increase in output each year and a particularly notable growth toward 2020. This suggests an overall expansion of research activity within the field during the observed window.

Regarding collaboration types, academic-only publications consistently account for the largest share of output across all years. Academic–industry collaborations represent a smaller but visible portion of total publications. The share of industrial collaboration, represented by the line in the figure, fluctuates over time rather than following a strictly monotonic trend. Aft

Overall, the figure indicates that while research productivity steadily increases over time, the proportion of industrial collaboration exhibits moderate variability. This suggests that growth in publication output does not automatically translate into a proportional increase in industry engagement.

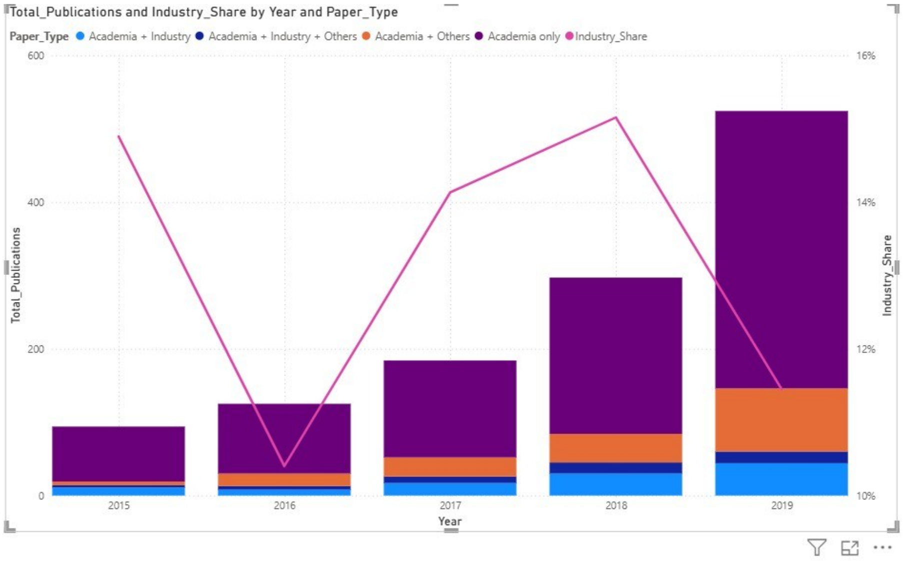


Figure 1. Publication output and industry collaboration share, 2015-2020

Figure 2 presents the temporal distribution of academic-only and academic–industry publications between 2015 and 2020. Across the entire period, the shares of both academic-only and academic–industry publications remain relatively stable, exhibiting only minor fluctuations from year to year. Academic-only publications consistently represent a substantially larger proportion of total output, clearly dominating the research landscape throughout the observed window.

Although the share of academic–industry collaboration varies slightly across years, these changes are modest and do not follow a consistent upward or downward trajectory. Despite small year-to-year variations in industry involvement, no clear monotonic or systematic trend is observed over the study period. Overall, the figure suggests structural stability in the composition of research output, with academic-only publications maintaining a persistently higher share compared to academic–industry collaborations.

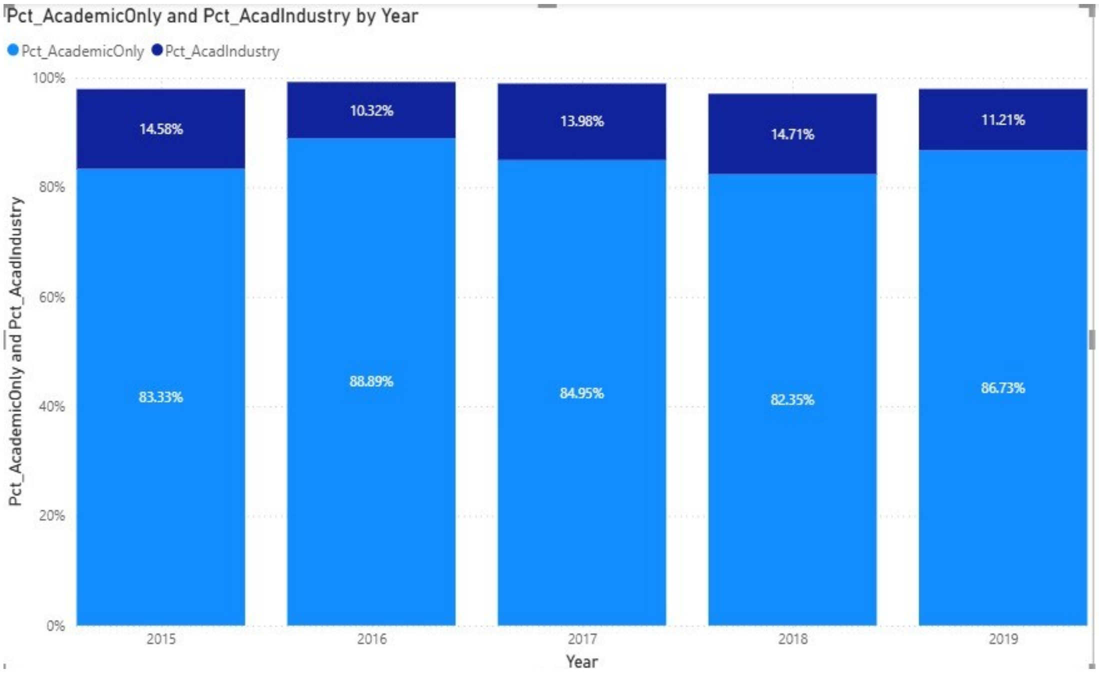


Figure 2. Share of academic and academic-industry publications by year

In figure 3, academic publications are divided into two groups based on their citation performance: those with citations below or equal to the mean, and those with citations above the mean. The bar chart compares the total number of publications in each group, while the line indicates the share of authors with industry affiliation relative to the total number of authors in each category. The results suggest that although highly cited publications are fewer in number, they exhibit a higher proportion of industry involvement, indicating that collaboration with industry may be associated with higher-impact scientific output, a pattern consistent with previous findings on the role of cross-sector collaboration in enhancing research visibility and impact (e.g., Perkmann et al., 2013).

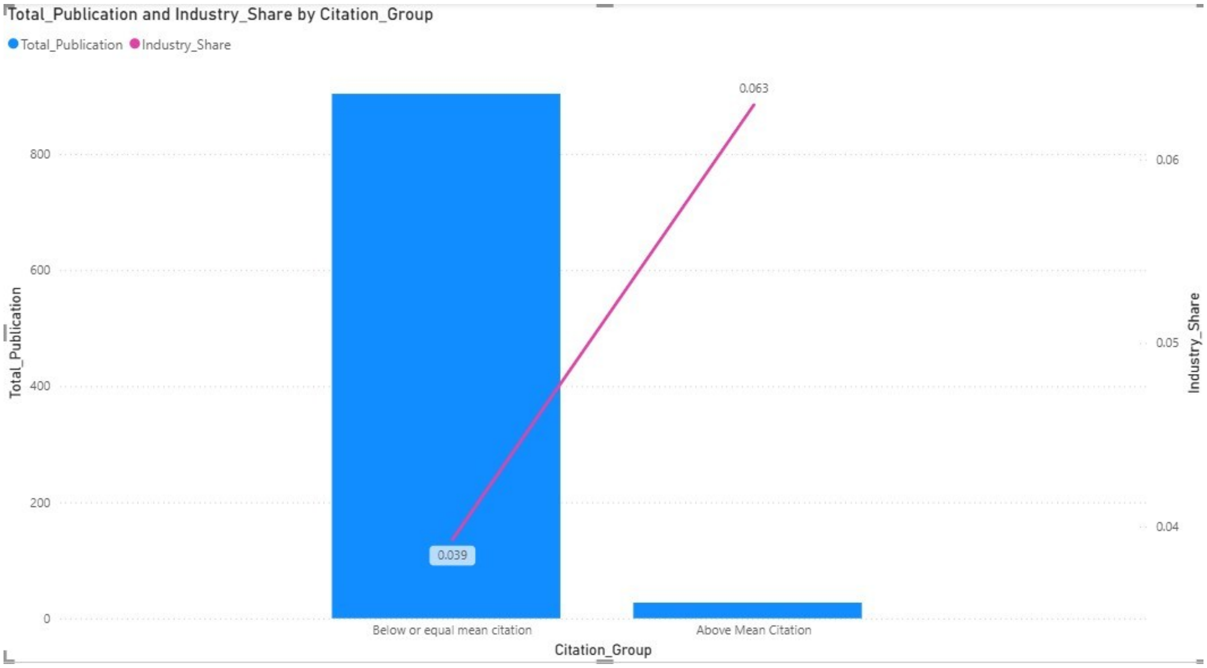


Figure 3. Publication volume and industry share by citation group

Figure 4 compares academic authors by dividing them into two groups based on their publication productivity, using the average number of publications as the threshold. The mean publication count is 1.21, which in practice corresponds to approximately one publication. Accordingly, authors are classified into those with publication counts below or equal to the average and those with above-average output. The bar chart shows the number of authors in each group, while the line represents the average share of industrial co-authors associated with each group. The results indicate that although highly productive authors are fewer in number, they tend to exhibit a higher level of collaboration with industry. This pattern suggests a positive association between research productivity and industrial engagement, potentially reflecting the role of industry partnerships in facilitating more resource-intensive and high-impact research activities (Perkmann et al., 2013)

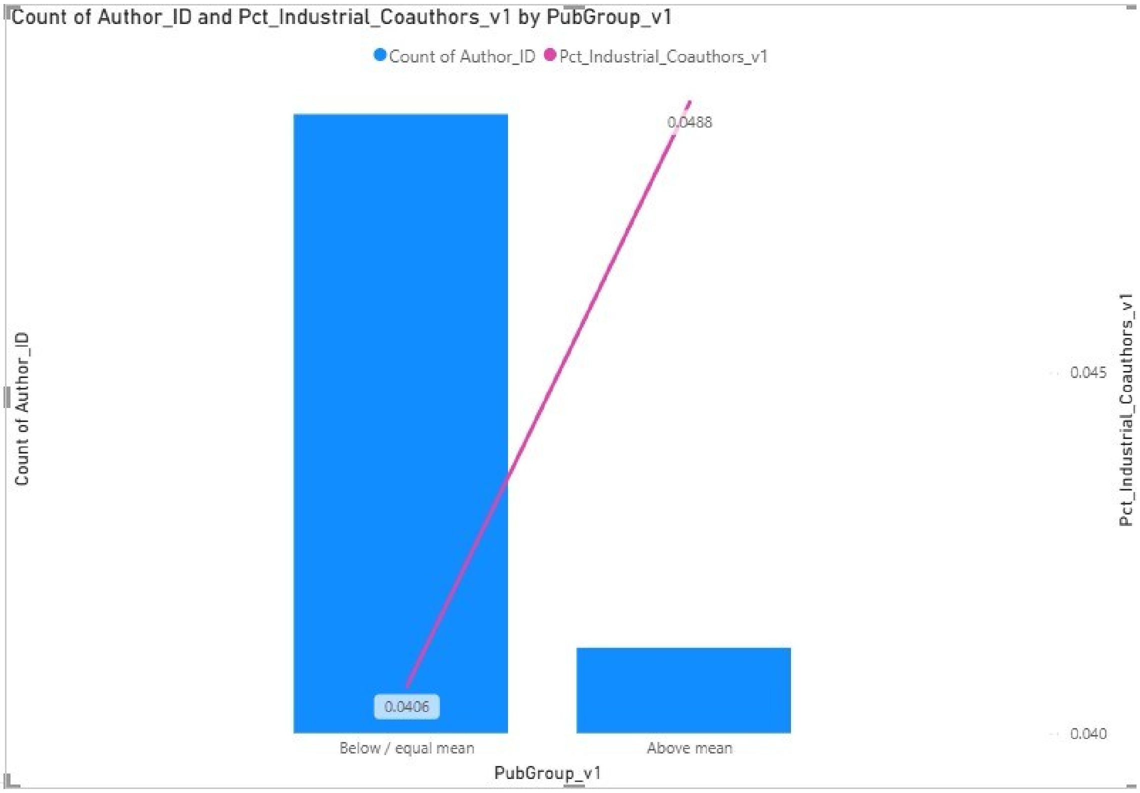


Figure 4. Industry collaboration across productivity groups

.

Figure 5. illustrates the average size of author teams across different types of publications, categorized by the composition of collaborating scientists. Specifically, it compares academic-only, industry-only, academic–industry, and other forms of collaboration in terms of the mean number of authors per paper. The results show that publications involving academic–industry collaboration have substantially larger team sizes compared to purely academic or industry-only papers, suggesting that cross-sector collaboration is associated with more complex research projects that require broader expertise and coordination. In contrast, industry-only publications tend to involve smaller teams, possibly reflecting more focused or proprietary research activities. This pattern is consistent with prior studies highlighting that inter-sector collaborations often lead to larger and more diverse research teams due to the integration of complementary knowledge and resources (e.g., Katz and Martin, 1997; Bozeman et al., 2013).

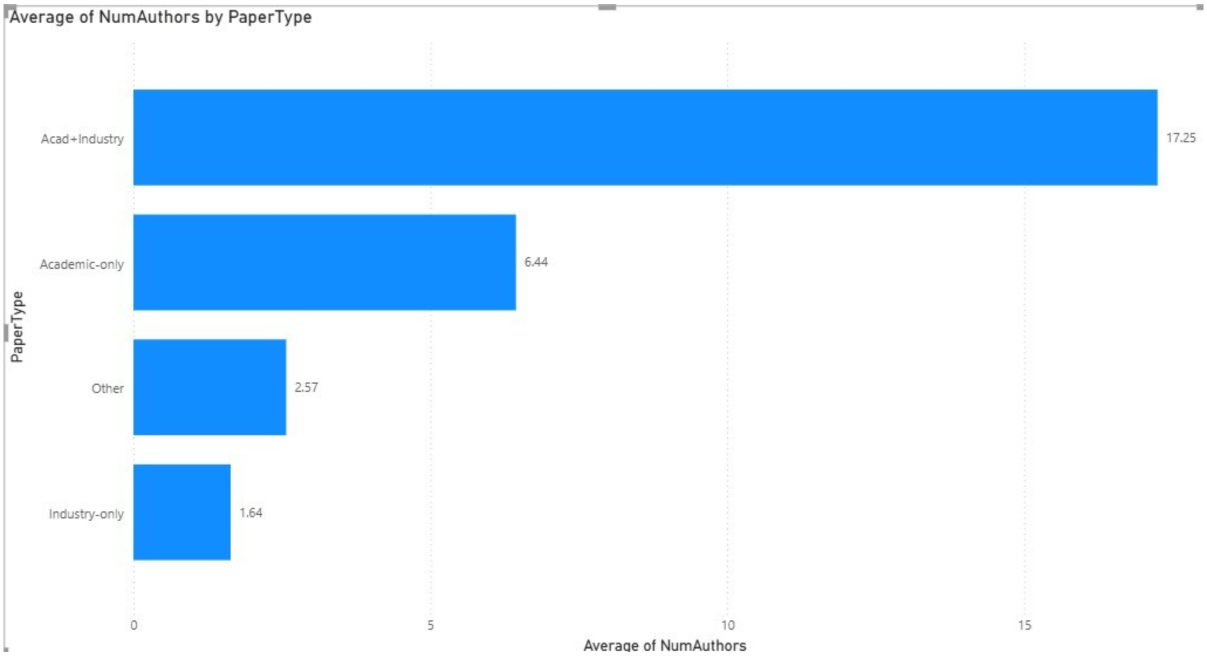


Figure 5. Average team size by paper type

Figure 6. displays the countries with which Canadian scientists have the highest levels of international collaboration, measured using a weighted count of collaborative publications. The chart shows that the United States is by far Canada’s most significant international research partner, followed at a considerable distance by China, Germany, the United Kingdom, and France. This pattern suggests that geographic proximity, strong institutional ties, shared research infrastructures, and linguistic and policy alignment play an important role in shaping international scientific collaboration. In particular, the dominant position of the United States reflects the close integration of the Canadian and U.S. research systems, especially in fast-growing and resource-intensive fields such as artificial intelligence (Katz and Martin, 1997; Wagner et al., 2015)

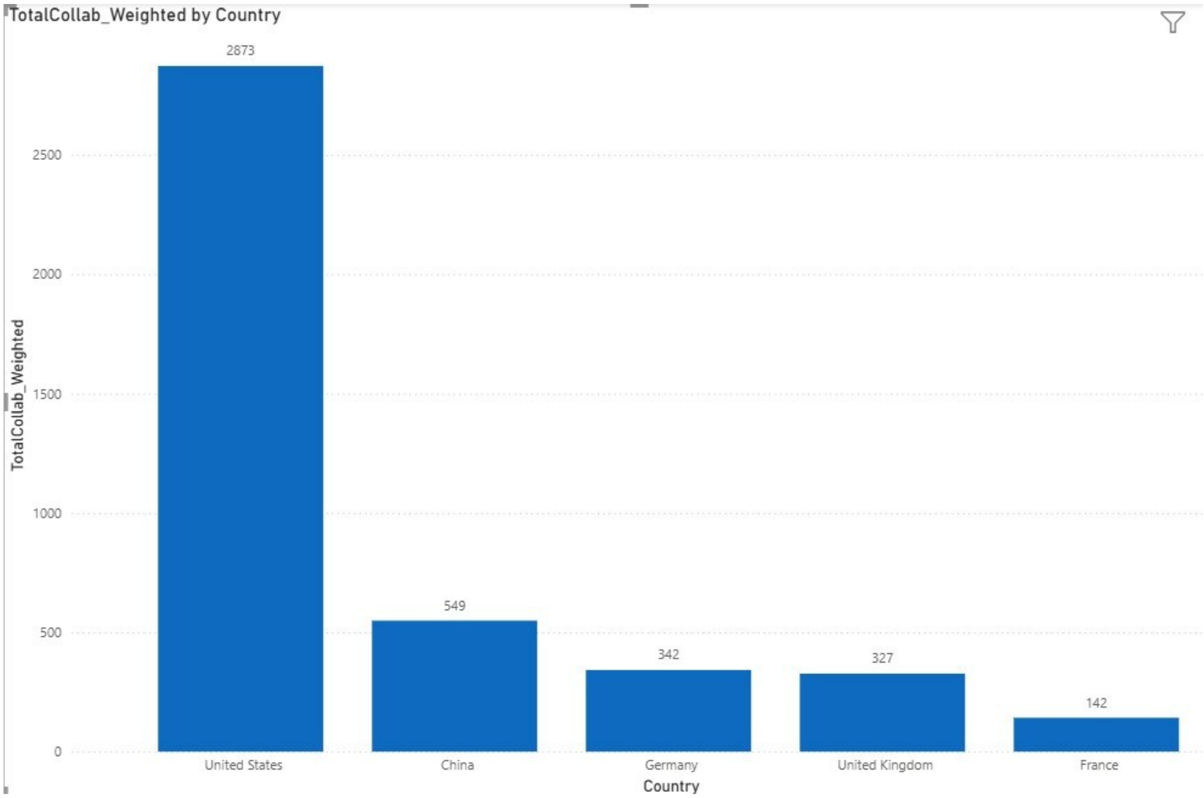


Figure 6. International collaboration by partner country

Figure 7 illustrates the yearly distribution of solo and collaborative publications between 2015 and 2020. The figure shows that collaborative publications consistently account for a larger share of total output than solo-authored works across all years. Notably, collaboration peaks around 2017, suggesting a temporary intensification of joint research efforts during this period. Overall, the figure indicates that collaboration constitutes the dominant mode of knowledge production, with no clear upward or downward trend over time. This pattern aligns with the broader literature emphasizing the growing importance and persistence of collaborative research practices in contemporary science, particularly in knowledge-intensive fields such as artificial intelligence (Wagner et al., 2015).

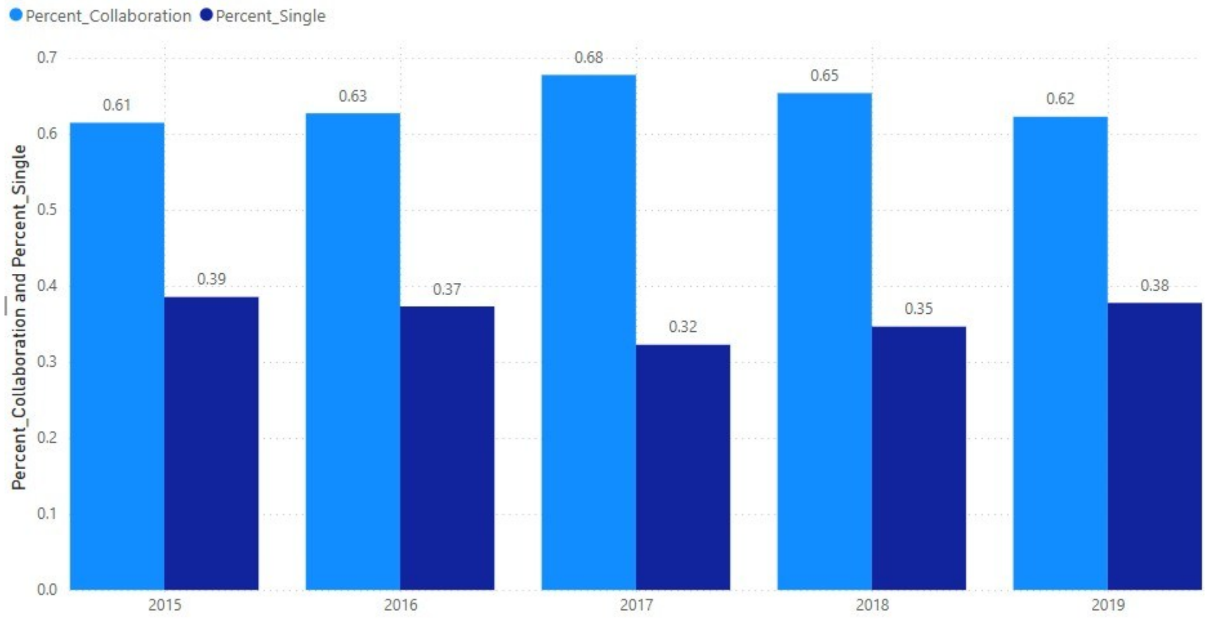


Figure 7. Collaborative and single-authored publications by year

Chapter 4: Statistical Analysis

This chapter presents the empirical framework used to examine how industry engagement and collaboration network structures influence the research performance of Canadian AI scientists. The analysis is conducted at the author level and is based on merged publication and patent data. Two dimensions of scientific performance are analyzed separately: publication productivity and citation impact. Given the distinct distributional properties of these outcomes, appropriate econometric models are employed for each. The following sections first define the variables used in the analysis and then present the empirical strategy and regression results.

4.1 Variables and Measurement

This section defines the dependent and independent variables employed in the empirical analysis. All variables are constructed at the author level based on merged Scopus publication data (2015–2020) and patent data from Lens. The dependent variables capture two dimensions of research performance, while the independent variables include collaboration characteristics, industry engagement measures, and structural network indicators.

4.1.1 Dependent Variables

4.1.1.1 Number of Publications

Publication productivity is measured as the total number of AI-related publications authored by each researcher during the 2015–2020 period. This variable counts the distinct publications attributed to each author in Scopus and captures the quantitative dimension of scientific performance at the author level. As a non-negative integer variable, it reflects the overall volume of research output produced by each scientist over the study window.

4.1.1.2 Mean Citation Impact

Citation impact is measured as the average number of citations per author over the 2015–2020 period. For each researcher, this variable is constructed by summing the citations received by all publications attributed to the author during the study window and dividing this total by the number of publications produced by the same author in that period. By normalizing total citations by publication count, this measure captures the average citation performance of a researcher’s body of work and distinguishes scientific impact from mere publication volume, thereby reflecting the relative influence of an author’s research rather than the sheer quantity of output.

4.1.2 Independent Variables

4.1.2.1Group Size

This indicator reflects the typical collaborative environment of a researcher. Higher values indicate systematic participation in larger research teams, whereas lower values suggest engagement in smaller groups or individual work.

4.1.2.2 Repeated Collaboration

Repeated collaboration reflects the extent to which a scientist’s collaborative behavior is characterized by sustained partnerships rather than one-time interactions.

4.1.2.3 Author-Inventor Status

Author–inventor status is a binary variable constructed at the author level. It takes the value of 1 only if a researcher is identified in both the Scopus publication dataset and the Lens patent dataset. This classification therefore indicates direct participation in both scientific research and technological innovation activities. Authors who appear only in the publication dataset and have no corresponding patent record in the Lens data are coded as 0.

4.1.2.4 Industry Participation

Percentage of industry collaboration (perc_indust) is constructed at the author level and measures the extent to which a researcher collaborates with industry-affiliated partners during the 2015–2020 period. The variable ranges from 0 to 100, and authors with no collaborators are assigned a value of zero.

4.1.2.5 Degree Centrality

Degree centrality is operationalized as the total number of unique co-authors directly connected to a given author in the network. The undirected co-authorship graph constructed for the 2015–2020 period was used as the input structure. In this graph, authors represent nodes and co-authorship ties represent edges. Higher values indicate a greater number of distinct collaborative partners.

4.1.2.6 Betweenness Centrality

Betweenness centrality (between) captures the extent to which a focal author lies on the shortest paths between other authors in the co-authorship network. Using the undirected co-authorship graph constructed for the 2015–2020 period, higher values indicate a greater bridging role across different collaboration groups within the network.

4.1.2.7 Team Embeddedness (Mean Cohesion)

Team embeddedness reflects how strongly an author's research teams are characterized by internal cohesion—i.e., the extent to which co-authors are connected to each other through other collaborations, rather than being linked only through the focal author.

4.2 Empirical Strategy

This study adopts an author-level regression framework to examine how collaboration patterns and network structure relate to the scientific performance of Canadian AI researchers during the 2015–2020 period. Two distinct dimensions of performance are analyzed to differentiate research quantity from research impact: publication productivity and citation impact.

Publication productivity is measured as the total number of publications per author during the observation window. Given that this outcome is a non-negative count variable, Poisson regression is employed. The model estimates the expected number of publications as a function of industry engagement (author–inventor status and percentage of industry collaboration), collaboration structure (average group size and repeated collaboration), and network position (degree centrality, betweenness centrality, and team embeddedness).

Citation impact is operationalized as the average number of citations per author, calculated by dividing total citations by publication count within the period. Because this measure is continuous and not a count variable, Ordinary Least Squares (OLS) regression is used. To address potential heteroskedasticity in cross-sectional data, heteroskedasticity-robust standard errors (HC3) are employed in all OLS specifications.

Prior to model estimation, all independent and dependent variables are examined for distributional properties. Where necessary, functional transformations are applied to ensure appropriate specification. To capture potential nonlinear relationships, quadratic terms are tested. In addition, interaction effects—particularly between industry engagement and network position—are evaluated to assess whether the impact of collaboration structure varies across different levels of industry involvement.

To mitigate multicollinearity, highly correlated network indicators are not included simultaneously in the same specification. Instead, alternative model versions are estimated to isolate the effects of different dimensions of collaboration and network structure. Model comparisons rely on standard goodness-of-fit criteria and robustness checks across specifications.

4.2.1 Sample Preparation and Data Screening

The analytical dataset consists of 3,083 unique Canadian AI authors identified from Scopus publications (2015–2020). All independent variables were examined for completeness, distributional characteristics, and potential data irregularities prior to model estimation. No missing observations were detected in these variables included in the regression analysis.

Descriptive analysis of the independent variables revealed substantial right-skewness and long-tailed distributions for several key network-related measures, including degree centrality, betweenness centrality, repeated collaboration intensity, industry collaboration intensity, and average group size. To stabilize scale differences, and better approximate linear relationships within the log-link regression framework, logarithmic transformations were applied to several of these variables.

Specifically, average group size was transformed using the natural logarithm (log) because this variable does not contain zero values. In contrast, several other independent variables—including degree centrality, betweenness centrality, percent repeat collaboration, and percent industry collaboration—contain zero values. For these variables, a log(1+x) transformation (log1p) was applied to preserve zero observations while reducing skewness and stabilizing variance.

In the case of betweenness centrality, the raw values were extremely small due to the normalization procedure used in constructing the collaboration network. To improve numerical stability and avoid precision issues during transformation, the betweenness values were first rescaled by multiplying them by 10^6 prior to applying the log(1+x) transformation. This rescaling does not change the relative ordering of observations but facilitates a more stable transformation and improves numerical interpretability within the regression models.

The team embeddedness mean variable also exhibits a right-skewed distribution. Despite this skewness, the variable was retained in its original scale for several reasons. Because team embeddedness is a bounded network measure ranging between 0 and 1, and logarithmic transformations would compress variation near zero and distort the substantive interpretation of the metric.

Finally, the percent industry collaboration variable exhibited a strongly zero-inflated and right-skewed distribution, with many authors reporting no industry engagement and a minority displaying substantial industry involvement. This pattern suggests a two-part structure combining participation and intensity. To capture these distinct dimensions, both a binary indicator of industry participation and the continuous percentage measure were retained.

4.3 Analysis of Publication Count

This section presents the empirical results for publication productivity using Poisson regression models. The estimation focuses on identifying how industry engagement, collaboration structure, and network position are associated with variations in the expected number of publications per author.

Model specifications are introduced progressively to evaluate the stability of effects across alternative configurations. Industry-related variables are first estimated, followed by the inclusion of collaboration characteristics and network centrality indicators. To avoid multicollinearity, highly correlated structural measures are not entered simultaneously; instead, alternative model versions are estimated to isolate distinct dimensions of network position.

Prior to final model selection, dispersion diagnostics were conducted to assess whether the variance of the dependent variable exceeded its mean. Given that overdispersion can lead to biased standard errors in Poisson models, a Negative Binomial (NB) specification was also considered as a robustness alternative. However, based on diagnostic evaluation and model comparison, the Poisson model was retained as the primary estimation strategy.

Quadratic terms are tested to capture potential nonlinear effects of collaboration variables. In addition, interaction terms are estimated to examine whether the relationship between network position and publication productivity varies conditional on industry engagement. Model fit is evaluated using log-likelihood values and information criteria (AIC) to compare competing specifications.

The dependent variable, the number of publications (n_pubs), represents a non-negative count outcome and exhibits substantial right-skewness. As shown in figure 8, the distribution of publications is highly concentrated at low values, wile many authors producing only one publication during the observation period, a small subset of authors reaches publication levels above four or five papers, resulting in a long right-tailed distribution. Despite this skewness, the dependent variable was retained in its original count form. Because the analytical framework relies on generalized linear models (GLM) with a log link function—specifically Poisson and Negative Binomial regressions—the distributional asymmetry of the dependent variable is explicitly modeled within the count-data framework. Transforming the dependent variable (e.g., using a logarithmic transformation) would obscure its discrete nature and complicate the interpretation of marginal effects. Therefore, the number of publications was modeled directly as a count outcome.

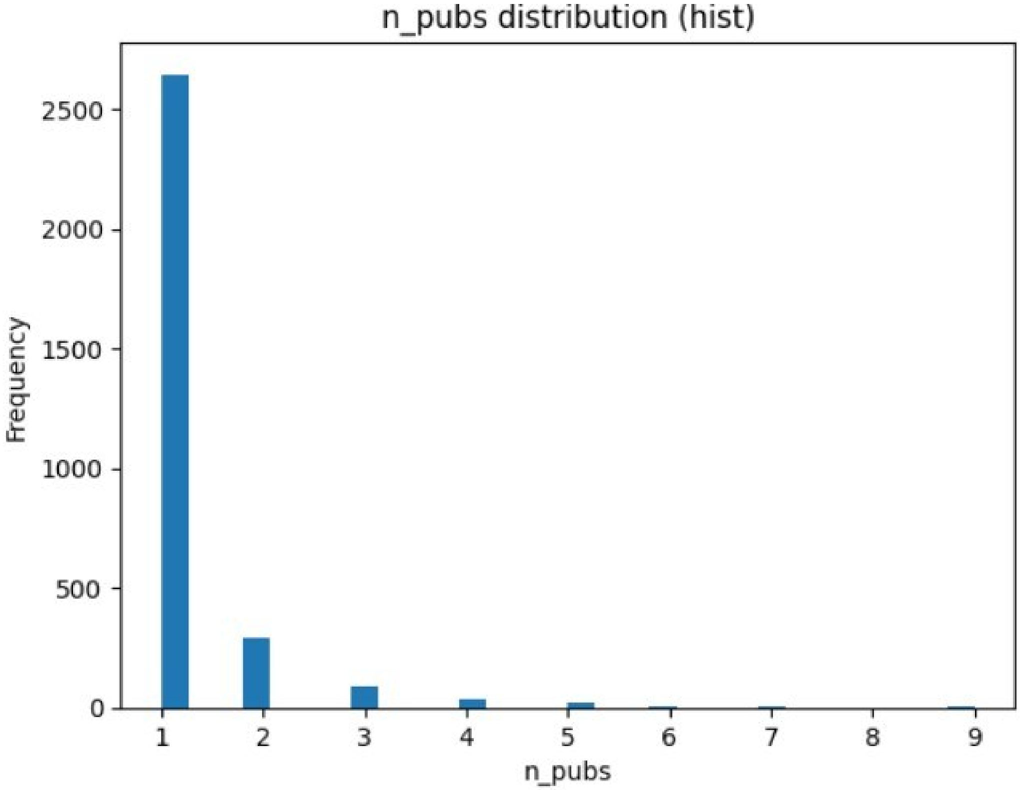


Figure 8. Frequency distribution of publication counts

4.3.1 Descriptive Statistics and Correlation Analysis

Given that this study employs the dependent variable of the total number of publications, assessing multicollinearity among the explanatory variables is essential prior to model estimation. The number of publications represents a non-negative count outcome and is therefore analyzed using count-data regression models (Poisson and Negative Binomial). Because multicollinearity can inflate standard errors and reduce the stability and interpretability of coefficient estimates in generalized linear models, it is necessary to evaluate the degree of correlation among the independent variables before estimating the regression models.

To assess potential collinearity, pairwise Pearson correlation coefficients were computed among all independent variables (Table 1). The results reveal two pairs of variables with high correlations. First, degree centrality (log1p_degree) and average group size (log_av_gr_size) show a correlation coefficient of more than 0.7, indicating that both variables capture highly related structural dimensions of collaboration breadth. Second, the binary industry participation indicator (bin_indust) and the continuous percentage of industry collaboration (log1p_perc_indust) exhibit a correlation exceeding 0.90, reflecting that both measures represent different operationalizations of industry engagement.

Because these values indicate a high risk of multicollinearity, these strongly correlated variables were not included simultaneously in the same regression specification. Instead, they were separated across alternative model specifications. As a result, four model specifications were constructed to ensure that highly correlated predictors do not enter the same model.

Specifically, the models are structured as follows:

Model 1 includes the network breadth measure degree centrality (log1p_degree) together with the continuous measure of industry collaboration intensity (log1p_perc_indust), along with the remaining control variables: betweenness centrality(log1p_between), percentage of repeated collaboration (log1p_perc_repeat), team embeddedness cohesion (team_embed_mean), and status of author inventor (auth_invent).

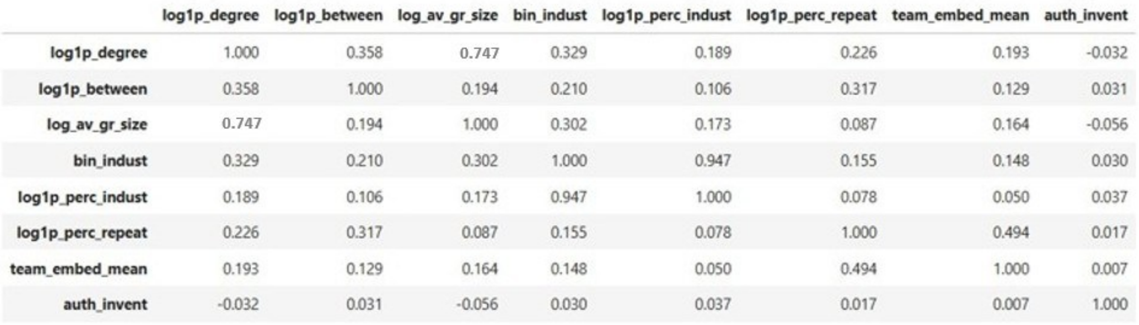
Model 2 includes degree centrality (log1p_degree) together with the binary industry participation indicator (bin_indust), while retaining the same set of control variables.

Model 3 replaces degree centrality with average group size (log_av_gr_size) while including the continuous measure of industry collaboration intensity (log1p_perc_indust), along with the remaining explanatory variables.

Model 4 includes average group size (log_av_gr_size) together with the binary industry participation indicator (bin_indust) and the same set of additional covariates.

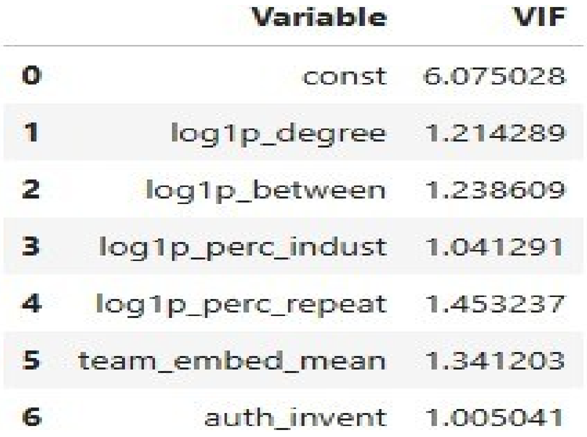
By separating the highly correlated predictors across these four specifications, the analysis avoids redundancy among explanatory variables and reduces the risk of inflated standard errors caused by multicollinearity. This modeling strategy allows the analysis to examine alternative operationalizations of collaboration breadth (degree centrality versus group size) and industry engagement (binary participation versus collaboration intensity) while maintaining statistical stability in both the count-data regression models for publication counts and the OLS models for citation outcomes.

Finally, to further validate the adequacy of the model structures, Variance Inflation Factors (VIF) were computed for each specification after separating the highly correlated structural variables. All VIF values remained well below conventional thresholds (VIF < 2), indicating no evidence of problematic multicollinearity in any of the estimated models.



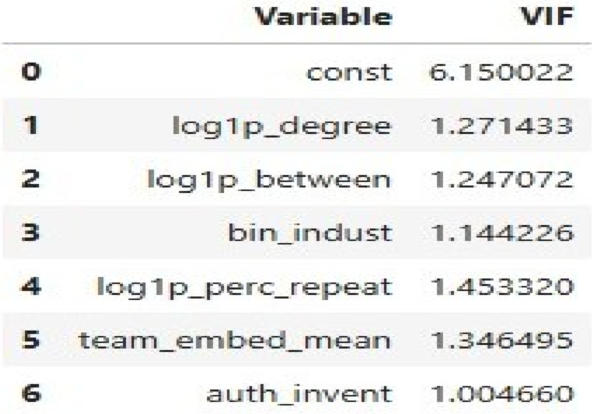
	log1p_degree	log1p_between	log_av_gr_size	bin_indust	log1p_perc_indust	log1p_perc_repeat	team_embed_mean	auth_invent
log1p_degree	1.000	0.358	0.747	0.329	0.189	0.226	0.193	-0.032
log1p_between	0.358	1.000	0.194	0.210	0.106	0.317	0.129	0.031
log_av_gr_size	0.747	0.194	1.000	0.302	0.173	0.087	0.164	-0.056
bin_indust	0.329	0.210	0.302	1.000	0.947	0.155	0.148	0.030
log1p_perc_indust	0.189	0.106	0.173	0.947	1.000	0.078	0.050	0.037
log1p_perc_repeat	0.226	0.317	0.087	0.155	0.078	1.000	0.494	0.017
team_embed_mean	0.193	0.129	0.164	0.148	0.050	0.494	1.000	0.007
auth_invent	-0.032	0.031	-0.056	0.030	0.037	0.017	0.007	1.000

Table 1. Pearson Correlation Matrix of Independent Variables



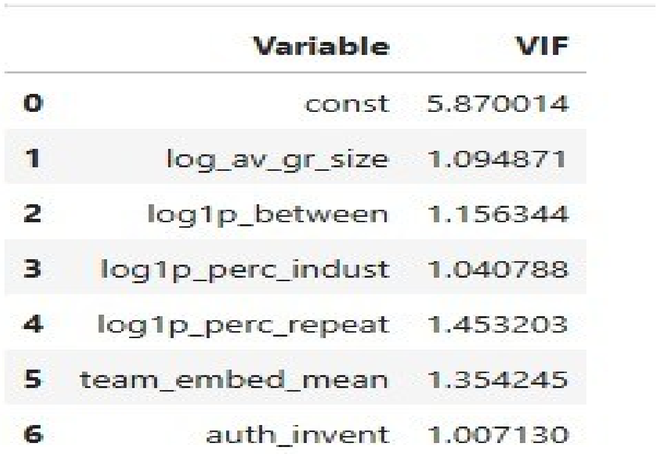
	Variable	VIF
0	const	6.075028
1	log1p_degree	1.214289
2	log1p_between	1.238609
3	log1p_perc_indust	1.041291
4	log1p_perc_repeat	1.453237
5	team_embed_mean	1.341203
6	auth_invent	1.005041

Table 2. Variance Inflation Factors (VIF)- Model 1



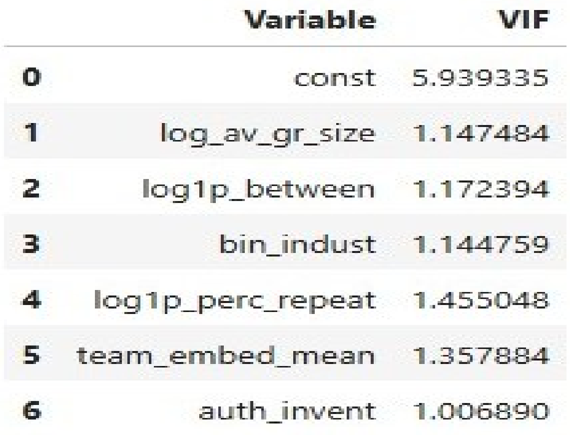
	Variable	VIF
0	const	6.150022
1	log1p_degree	1.271433
2	log1p_between	1.247072
3	bin_indust	1.144226
4	log1p_perc_repeat	1.453320
5	team_embed_mean	1.346495
6	auth_invent	1.004660

Table 3. Variance Inflation Factors (VIF)- Model 2



	Variable	VIF
0	const	5.870014
1	log_av_gr_size	1.094871
2	log1p_between	1.156344
3	log1p_perc_indust	1.040788
4	log1p_perc_repeat	1.453203
5	team_embed_mean	1.354245
6	auth_invent	1.007130

Table 4. Variance Inflation Factors (VIF)- Model 3



	Variable	VIF
0	const	5.939335
1	log_av_gr_size	1.147484
2	log1p_between	1.172394
3	bin_indust	1.144759
4	log1p_perc_repeat	1.455048
5	team_embed_mean	1.357884
6	auth_invent	1.006890

Table 5. Variance Inflation Factors (VIF)- Model 4

4.3.2 Model Selection

The dependent variable in this study is the number of publications (n_pubs) produced by each author during the observation period. Because this variable represents a non-negative integer count outcome, regression models designed for count data are required. Two standard approaches are commonly applied in such contexts: Poisson regression and Negative Binomial regression, both belonging to the generalized linear model (GLM) framework with a log-link function.

To determine the most appropriate specification, the distributional properties of the dependent variable were first examined. Descriptive statistics indicate that publication counts are highly concentrated at low values. The mean number of publications in the sample is approximately 1.22, while the variance is approximately 0.42, indicating that the variance is substantially lower than the mean. This pattern suggests the presence of underdispersion rather than overdispersion in the publication data.

To further evaluate the dispersion structure, a baseline Poisson model was estimated, and the Pearson dispersion statistic was computed. The resulting value (0.09) is well below unity, providing strong evidence of underdispersion. In count-data modeling, the Negative Binomial model is primarily designed to address overdispersion, where the conditional variance exceeds the conditional mean. Because the publication data in this study do not exhibit overdispersion, the Negative Binomial specification does not provide a theoretical advantage.

Model comparison based on the Akaike Information Criterion (AIC) further supports this conclusion. The Poisson model yields an AIC value of 6758, whereas the Negative Binomial model produces a substantially higher AIC of 9231, indicating a considerably poorer fit. Taken together, the dispersion diagnostics and information-criterion comparison consistently indicate that the Poisson regression model provides the more appropriate specification for modeling publication counts in this study.

Although the Poisson model assumes equality between the conditional mean and variance, empirical data may still exhibit minor deviations from this assumption. To ensure reliable statistical inference and mitigate potential effects of dispersion misspecification, heteroskedasticity-robust standard errors (HC3) were applied to all Poisson regression estimations.

4.3.3 Theoretical Model Specification

Following the selection of the Poisson count model, the specification of explanatory variables was guided by theoretical insights from the literature on scientific collaboration networks and knowledge exchange between academia and industry. Prior research suggests that both network positions within collaboration structures and different forms of industry engagement can shape the research productivity of scientists.

First, variables capturing collaboration structure within co-authorship networks were incorporated. The average group size represents the typical number of collaborators involved in a scientist's publications and reflects the breadth of research collaboration. While larger teams may provide access to complementary expertise, diverse knowledge resources, and research infrastructure, the literature also highlights potential coordination costs associated with large collaborative teams.

Another key network measure included in the analysis is betweenness centrality, which captures the extent to which a scientist occupies a brokerage position within the co-authorship network. Researchers with high betweenness centrality often serve as bridges connecting otherwise disconnected research groups, allowing them to access diverse knowledge flows and potentially enhance innovative output. At the same time, occupying brokerage positions may involve additional coordination demands and communication costs. In addition to brokerage positions, the analysis also considers the overall connectivity of researchers within the collaboration network. Degree centrality captures the number of direct collaborations ties a scientist maintains within the co-authorship network and reflects the breadth of their collaborative relationships. Researchers with higher degree centrality typically have access to a larger set of collaborators, which may increase opportunities for knowledge exchange, resource sharing, and joint research output. Prior studies in the science of science literature suggest that highly connected researchers often benefit from broader access to information and collaborative opportunities, potentially enhancing their scientific productivity.

Second, the model incorporates two distinct forms of industry engagement. The first is the percentage of industry collaborators (perc_indust), which measures the share of a scientist's collaborators affiliated with industry organizations. Because this variable contains a large proportion of zero values, indicating that many scientists do not collaborate with industry at all, an alternative binary specification (bin_indust) was also constructed. These binary variable captures whether a researcher engages in industry collaboration at least once, allowing the analysis to distinguish between the presence of industry collaboration and the intensity of such collaboration. The second indicator of industry engagement is the author–inventor status (auth_invent). This variable identifies researchers who are simultaneously involved in scientific publication and patenting activities, reflecting a more direct linkage between academic research and technological innovation. In the context of university–industry knowledge transfer, author–inventors are often considered key actors who bridge the scientific and technological domains.

Finally, repeated collaboration (perc_repeat) and team embeddedness (team_embed_mean) are included as control variables capturing additional dimensions of collaboration structure. Repeated collaboration reflects the stability of research partnerships, while team embeddedness measures the extent to which collaborators are interconnected within research teams, potentially influencing knowledge exchange efficiency.

In addition to the main effects, the analysis considers potential nonlinear and interaction relationships motivated by theoretical arguments in the collaboration and innovation literature.

First, a quadratic specification of betweenness centrality is introduced to test for potential nonlinear effects of brokerage positions on research productivity. Moderate levels of brokerage can facilitate access to diverse information sources and enhance knowledge recombination. However, extremely high brokerage positions may increase coordination burdens, information overload, and communication costs. In large collaboration networks, scientists occupying very central brokerage roles may face constraints in managing multiple collaborative ties simultaneously, which can reduce their effective research productivity. These mechanisms suggest that the relationship between brokerage position and publication output may exhibit diminishing returns, making the quadratic specification theoretically meaningful.

In addition, a quadratic specification of average group size is examined to capture potential nonlinear effects of collaboration breadth. A growing body of research on scientific collaboration suggests that the relationship between team size and research productivity may not be strictly linear. While larger teams can benefit from a broader pool of expertise, complementary skills, and access to diverse knowledge resources, increasing team size can also generate coordination costs, communication overhead, and managerial complexity. As the number of collaborators grows, coordination among team members becomes more difficult, potentially reducing the marginal productivity of additional collaborators. Several studies highlight this tension between the benefits and costs of collaboration. On the one hand, larger teams may facilitate knowledge recombination and interdisciplinary problem solving (Wuchty et al., 2007). On the other hand, excessively large teams may suffer from coordination inefficiencies and reduced individual contribution, which can negatively affect research output (Cummings & Kiesler, 2005). Similarly, Jones et al. (2008) argue that increasing collaboration size may lead to higher coordination demands and slower knowledge integration processes. These mechanisms suggest that the relationship between team size and scientific productivity may exhibit diminishing returns, making it theoretically appropriate to test for a nonlinear (quadratic) effect of group size in the regression models.

Second, the models examine interaction effects between collaboration structure and industry engagement. One important interaction considered is between industry collaboration and group size, reflecting the idea that the productivity implications of industry partnerships may depend on the size of the research team. Industry collaboration within larger teams may introduce additional coordination complexity, potentially moderating the positive effects of collaboration on publication output.

Furthermore, the analysis also examines the interaction between author–inventor status and betweenness centrality. Scientists who simultaneously engage in academic publishing and patenting activities often operate at the intersection of scientific and technological knowledge networks. When such researchers also occupy brokerage positions within collaboration networks, they may gain privileged access to heterogeneous knowledge sources that facilitate innovation and knowledge recombination.

To account for these theoretical considerations, several alternative model specifications are estimated. Baseline models include the primary explanatory variables, while extended models incorporate nonlinear and interaction terms. This strategy allows the analysis to evaluate whether these theoretically motivated mechanisms improve the explanatory power of the models.

4.3.4 Baseline Models

Following the theoretical specification of explanatory variables discussed in the previous section, a set of baseline regression models was constructed to examine the relationship between collaboration structures, industry engagement, and scientific productivity. The dependent variable in these models is the number of publications produced by each author (n_pubs) during the observation period. Because publication counts represent non-negative integer outcomes, the models are estimated using a Poisson regression framework with a log-link function.

Formally, the expected number of publications for author i is modeled as:

$$E(n\_pubs_i|X_i) = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki})$$

where X_i represents the vector of explanatory variables capturing collaboration structure and industry engagement.

As discussed earlier in the correlation analysis, some explanatory variables exhibit strong associations. Betweenness centrality and group size show a high level of correlation, while the two industry indicators are also strongly related. To avoid potential multicollinearity and ensure stable coefficient estimation, these highly correlated variables are introduced separately across different model specifications.

Accordingly, four baseline Poisson models are estimated. These models share a common structure but differ in the inclusion of the highly correlated variables representing collaboration structure and industry engagement.

Estimating multiple baseline specifications allows the analysis to examine whether the effects of collaboration structures and industry engagement remain consistent across alternative operationalizations of these variables. These baseline models also provide the foundation for the subsequent analysis, in which nonlinear (quadratic) terms and interaction effects are introduced to further explore the mechanisms linking collaboration structures, industry engagement, and publication productivity.

4.3.5 Baseline Models Estimation and Comparison

Tables 6–9 present the estimation results of the four baseline Poisson regression models examining the relationship between collaboration structures, industry engagement, and scientific productivity. All models are estimated using Poisson generalized linear models with a log-link function and robust standard errors.

Across the model specifications, several variables show consistent and statistically significant associations with publication output. Betweenness centrality exhibits a positive and highly significant relationship with the number of publications across all models. This result suggests that researchers occupying brokerage positions within the co-authorship network tend to produce more scientific publications, likely due to their access to diverse knowledge flows and collaborative opportunities (Burt, 2004; Newman, 2001).

The results also indicate a negative and statistically significant effect of group size in the models where this variable is included. This finding suggests that larger collaboration teams may be associated with lower individual publication productivity, potentially reflecting coordination costs, communication complexity, and increasing managerial challenges in large research groups (Cummings & Kiesler, 2005; Wuchty et al., 2007).

In contrast, degree centrality does not show a statistically significant effect on publication counts in the models where it is included. It is important to note that degree centrality is strongly correlated with group size in co-authorship networks, since researchers who frequently participate in larger teams often accumulate a larger number of collaborators across publications. However, the two measures capture different dimensions of collaboration. Degree centrality reflects the number of distinct collaborators a scientist connects with across the network, whereas group size reflects the number of co-authors involved in individual publications. As a result, a researcher may repeatedly publish within relatively large but stable teams—leading to a large group size but relatively limited diversity of collaborators—while another researcher may collaborate with many different partners across multiple publications through smaller teams, resulting in high degree centrality. Because these two measures represent network reach versus team composition, they may generate different productivity implications (Newman, 2001; Uzzi & Spiro, 2005). Consequently, even when these variables are statistically correlated, they may produce different effects in regression models of scientific productivity.

Regarding industry engagement, the results indicate that the author–inventor indicator is positively and statistically significant, suggesting that researchers who are simultaneously involved in scientific publishing and patenting activities tend to exhibit higher scientific productivity. In the literature on university–industry knowledge transfer, such individuals are often described as academic inventors who operate at the interface between scientific research and technological innovation. These researchers frequently act as intermediaries linking academic knowledge production with industrial application, enabling more efficient knowledge flows between the scientific and technological domains (Cohen et al., 2002; Graham & Sampat, 2007). Because author–inventors participate in both publication and patenting activities, they often benefit from broader knowledge exposure, access to industrial research problems, and stronger collaboration networks, which can enhance scientific productivity. Previous studies have also shown that academic inventors tend to publish more and produce research with greater scientific impact compared with purely academic researchers (Stephan, 2012; Harhoff & Hoisl, 2007). The positive coefficient observed for the author–inventor variable in the regression models is therefore consistent with these findings, indicating that stronger integration between academic research and technological innovation can be associated with higher levels of scientific publication output. In contrast, the direct measures of industry collaboration (perc_indust and bin_indust) do not appear to have statistically significant effects in the baseline specifications. This result suggests that the direct effect of industry collaboration on publication productivity may be limited or context dependent. Previous empirical studies have similarly documented mixed evidence regarding the relationship between industry collaboration and scientific publishing. While some studies find positive effects, others report neutral or even negative relationships, particularly when industrial collaboration shifts research activities toward applied technological outcomes or patenting rather than academic dissemination (Stephan, 2012; Perkmann et al., 2013). Therefore, the absence of a statistically significant direct effect for these variables in the baseline models does not necessarily imply that industry collaboration is irrelevant for scientific productivity. Instead, it suggests that the influence of industry engagement may operate through more complex mechanisms, such as interactions with collaboration structures or nonlinear relationships within research networks. For example, the productivity implications of industry collaboration may depend on the size of research teams or on the position of researchers within collaboration networks. These mechanisms are explored in the extended model specifications that incorporate quadratic terms and interaction effects in the subsequent analysis.

Among the control variables, repeated collaboration (perc_repeat) shows a strong and positive association with publication output across all models, indicating that stable collaborative relationships are associated with higher research productivity. Prior studies suggest that repeated collaborations may improve coordination and trust among collaborators, facilitating more efficient knowledge exchange (Uzzi & Spiro, 2005).

Conversely, team embeddedness exhibits a negative and statistically significant relationship with publication counts, suggesting that highly cohesive collaboration structures may limit exposure to diverse knowledge sources and reduce opportunities for novel knowledge recombination.

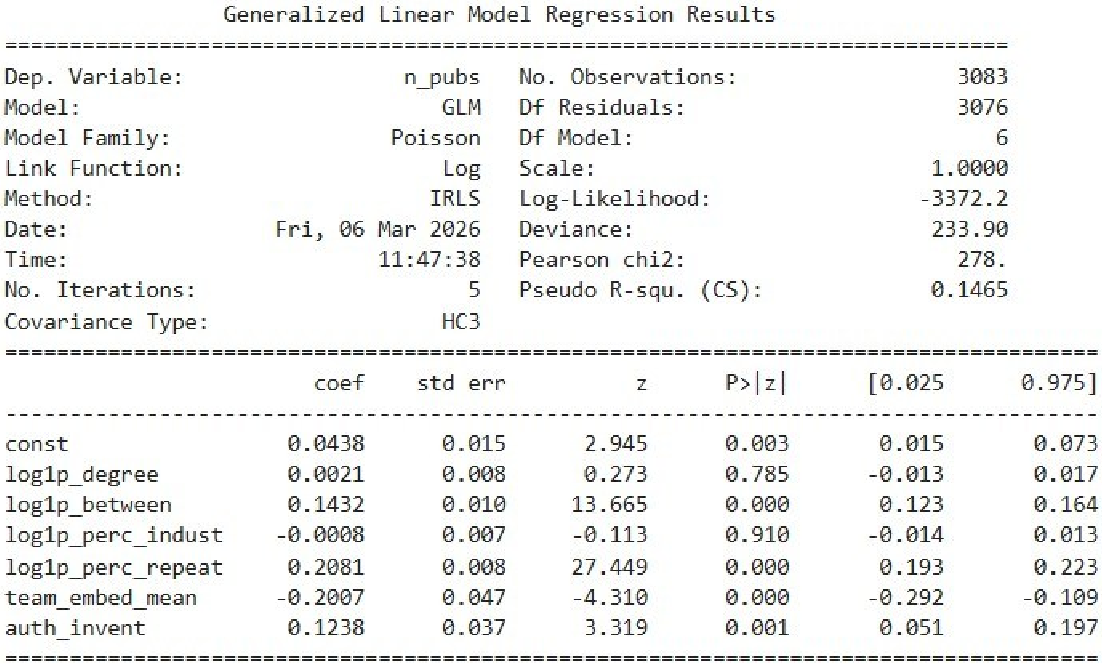
Table 10 presents a comparison of the goodness-of-fit statistics for the four baseline Poisson regression models examining the relationship between collaboration structures, industry engagement, and scientific productivity. Model performance is evaluated using several commonly used criteria in count data modeling, including the log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Pearson chi-square statistics. These measures are widely used for model comparison in generalized linear models, where lower AIC and BIC values and higher log-likelihood values indicate better model fit (Akaike, 1974; Burnham & Anderson, 2002).

The results show that the models incorporating group size as the primary collaboration structure variable (Models 3 and 4) outperform those including degree centrality (Models 1 and 2) across all model fit criteria. Model 3 exhibits the highest log-likelihood value (−3368.115) compared with −3372.153 and −3372.000 for Models 1 and 2, respectively. Similarly, the AIC value for Model 3 (6750.230) is substantially lower than the AIC values of the degree-based models (6758.307 for Model 1 and 6758.000 for Model 2), indicating a clear improvement in model fit. The magnitude of the AIC differences further confirms this result. The difference in AIC between Model 3 and the degree-based models exceeds 8 units, which represents strong evidence in favor of Model 3 according to established model selection guidelines (Burnham & Anderson, 2002). By contrast, the AIC difference between Model 3 and Model 4 is extremely small (ΔAIC $\approx$ 0.06), indicating that the two group-size specifications perform almost equivalently in terms of explanatory power.

The comparison of BIC values leads to a similar conclusion. Model 3 produces the lowest BIC value (−24485.709) among all models, followed closely by Model 4 (−24485.648), while the degree-based models show substantially higher BIC values. Since BIC penalizes model complexity more strongly than AIC, this result further supports the superiority of the group-size specifications.

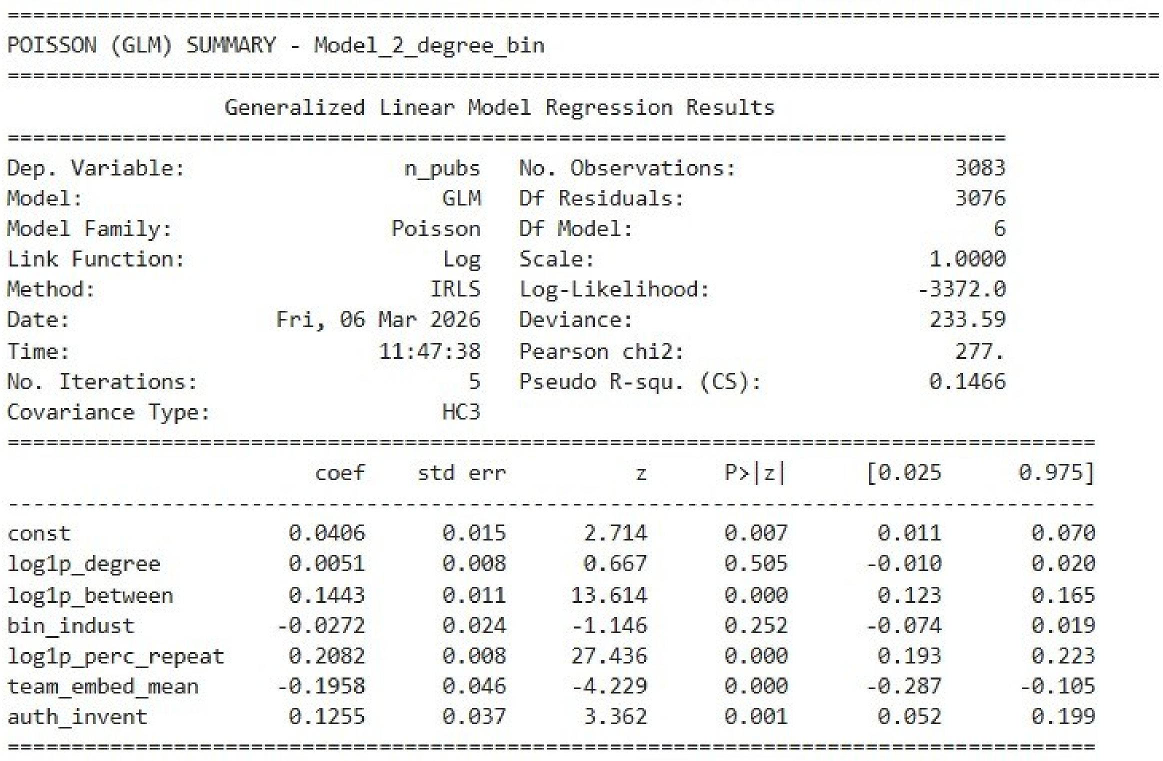
A similar pattern is observed when comparing the Pearson chi-square statistics, which are commonly used to evaluate the goodness of fit of Poisson models. Model 3 yields the lowest Pearson chi-square value (265.065), slightly outperforming Model 4 (265.150), while Models 1 and 2 show substantially higher values (277.580 and 277.163, respectively). Lower Pearson chi-square values indicate a better correspondence between the observed and predicted values of the dependent variable.

Taken together, these results consistently indicate that Models 3 and 4 provide a substantially better fit to the data than the degree-based specifications. Among the four models, Model 3 provides the best overall performance, as it combines the highest log-likelihood with the lowest AIC, BIC, and Pearson chi-square statistics. Therefore, Model 3 is selected as the preferred baseline specification for the subsequent analysis. The results also suggest that team size captures collaboration structure more effectively than degree centrality in explaining variation in publication productivity within the co-authorship network.



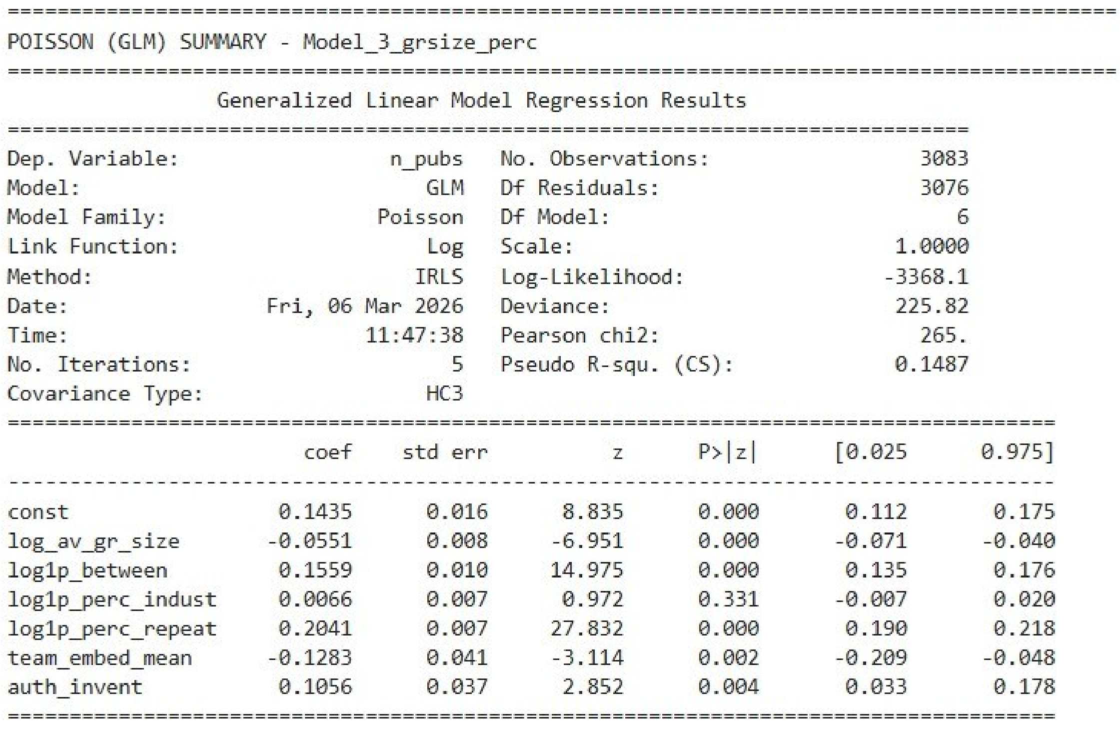
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3076
Model Family:                 Poisson   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3372.2
Date:                Fri, 06 Mar 2026   Deviance:                       233.90
Time:                        11:47:38   Pearson chi2:                     278.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1465
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.0438      0.015      2.945      0.003       0.015       0.073
log1p_degree          0.0021      0.008      0.273      0.785      -0.013       0.017
log1p_between         0.1432      0.010     13.665      0.000       0.123       0.164
log1p_perc_indust    -0.0008      0.007     -0.113      0.910      -0.014       0.013
log1p_perc_repeat     0.2081      0.008     27.449      0.000       0.193       0.223
team_embed_mean      -0.2007      0.047     -4.310      0.000      -0.292      -0.109
auth_invent           0.1238      0.037      3.319      0.001       0.051       0.197
=====================================================================================
```

Table 6. Poisson regression results for publications-M1



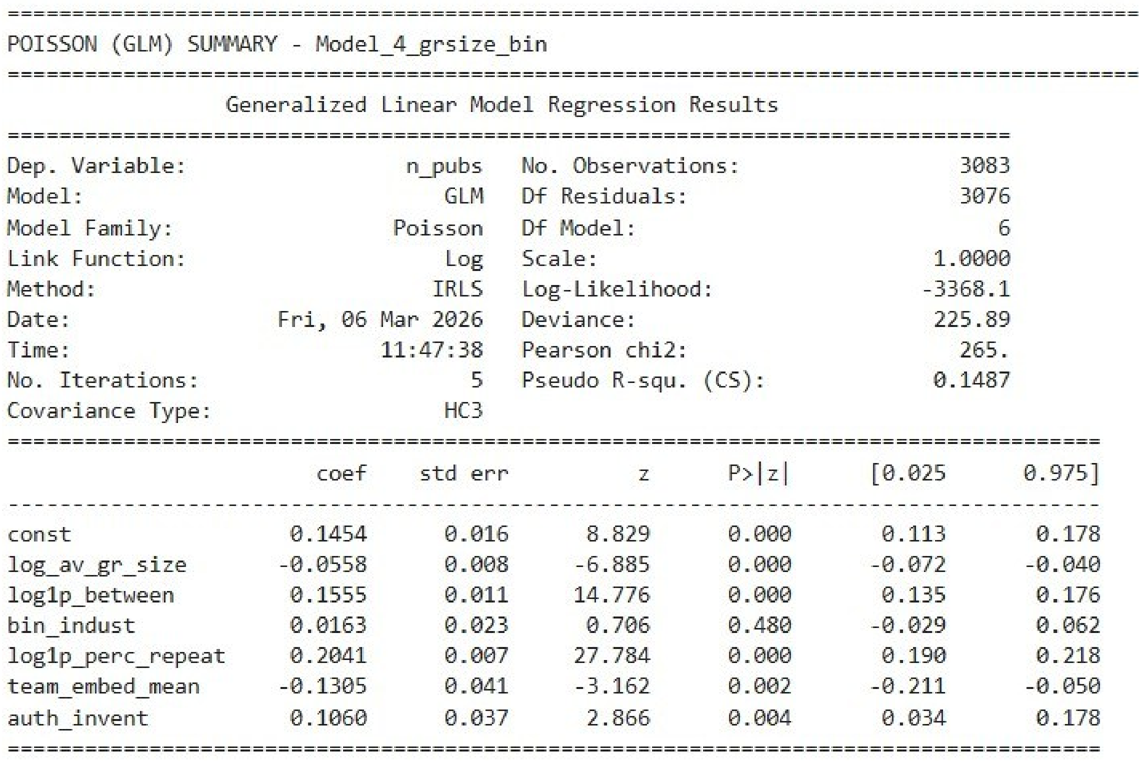
```
==========================================================================================
POISSON (GLM) SUMMARY - Model_2_degree_bin
==========================================================================================
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3076
Model Family:                 Poisson   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3372.0
Date:                Fri, 06 Mar 2026   Deviance:                       233.59
Time:                        11:47:38   Pearson chi2:                     277.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1466
Covariance Type:                  HC3                                         
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.0406      0.015      2.714      0.007       0.011       0.070
log1p_degree          0.0051      0.008      0.667      0.505      -0.010       0.020
log1p_between         0.1443      0.011     13.614      0.000       0.123       0.165
bin_indust           -0.0272      0.024     -1.146      0.252      -0.074       0.019
log1p_perc_repeat     0.2082      0.008     27.436      0.000       0.193       0.223
team_embed_mean      -0.1958      0.046     -4.229      0.000      -0.287      -0.105
auth_invent           0.1255      0.037      3.362      0.001       0.052       0.199
=====================================================================================
```

Table 7. Poisson regression results for publications-M2



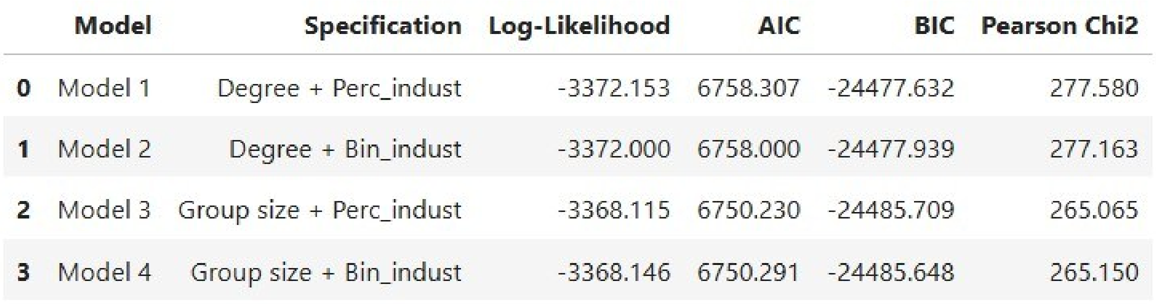
```
==========================================================================================
POISSON (GLM) SUMMARY - Model_3_grsize_perc
==========================================================================================
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3076
Model Family:                 Poisson   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3368.1
Date:                Fri, 06 Mar 2026   Deviance:                       225.82
Time:                        11:47:38   Pearson chi2:                     265.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1487
Covariance Type:                  HC3                                         
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1435      0.016      8.835      0.000       0.112       0.175
log_av_gr_size       -0.0551      0.008     -6.951      0.000      -0.071      -0.040
log1p_between         0.1559      0.010     14.975      0.000       0.135       0.176
log1p_perc_indust     0.0066      0.007      0.972      0.331      -0.007       0.020
log1p_perc_repeat     0.2041      0.007     27.832      0.000       0.190       0.218
team_embed_mean      -0.1283      0.041     -3.114      0.002      -0.209      -0.048
auth_invent           0.1056      0.037      2.852      0.004       0.033       0.178
=====================================================================================
```

Table 8. Poisson regression results for publications-M3



```
==================================================================================
POISSON (GLM) SUMMARY - Model_4_grsize_bin
==================================================================================
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3076
Model Family:                 Poisson   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3368.1
Date:                Fri, 06 Mar 2026   Deviance:                       225.89
Time:                        11:47:38   Pearson chi2:                     265.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1487
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1454      0.016      8.829      0.000       0.113       0.178
log_av_gr_size       -0.0558      0.008     -6.885      0.000      -0.072      -0.040
log1p_between         0.1555      0.011     14.776      0.000       0.135       0.176
bin_indust            0.0163      0.023      0.706      0.480      -0.029       0.062
log1p_perc_repeat     0.2041      0.007     27.784      0.000       0.190       0.218
team_embed_mean      -0.1305      0.041     -3.162      0.002      -0.211      -0.050
auth_invent           0.1060      0.037      2.866      0.004       0.034       0.178
=====================================================================================
```

Table 9. Poisson regression results for publications-M4



	Model	Specification	Log-Likelihood	AIC	BIC	Pearson Chi2
0	Model 1	Degree + Perc_indust	-3372.153	6758.307	-24477.632	277.580
1	Model 2	Degree + Bin_indust	-3372.000	6758.000	-24477.939	277.163
2	Model 3	Group size + Perc_indust	-3368.115	6750.230	-24485.709	265.065
3	Model 4	Group size + Bin_indust	-3368.146	6750.291	-24485.648	265.150

Table 10. Model fit comparison for baseline Poisson models

4.3.6 Extended Model Specifications: Nonlinear and Interaction Effects

To further examine whether the relationship between collaboration structures and scientific productivity follows nonlinear patterns, additional Poisson regression models were estimated by introducing quadratic specifications for key network variables. Quadratic terms were included for betweenness centrality and group size, two structural indicators capturing different aspects of collaboration within the co-authorship network.

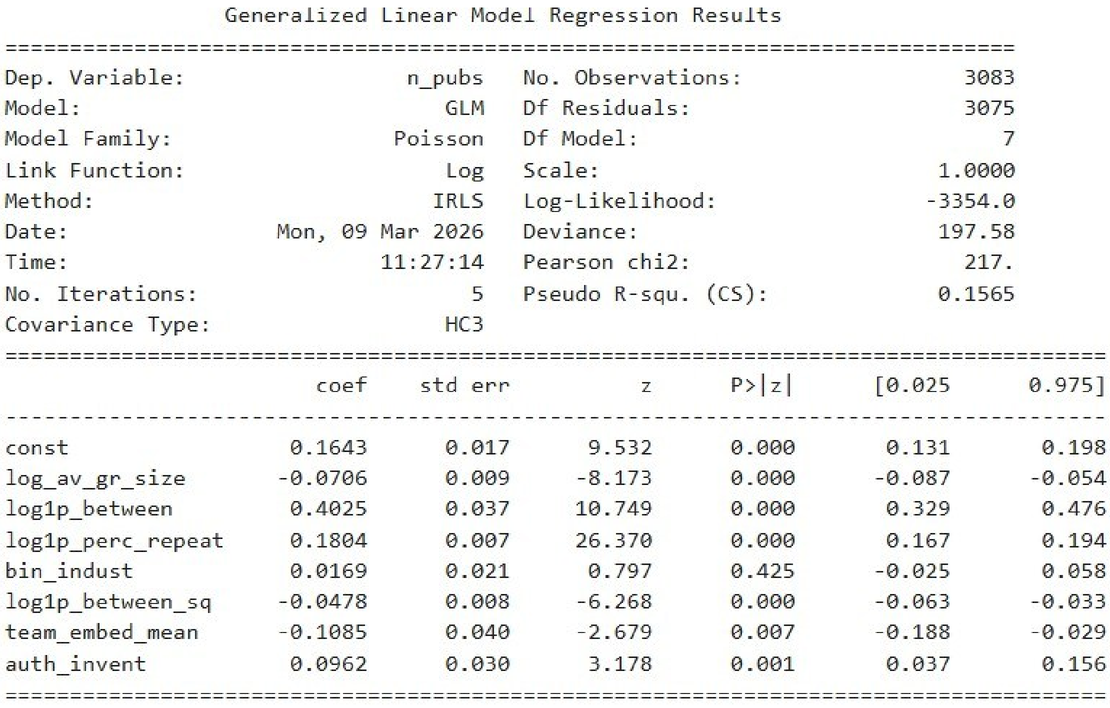
Table 11 presents the results of the model including a quadratic specification of betweenness centrality. The results indicate that the linear term of betweenness centrality is positive and statistically significant ($\beta = 0.4025$, $p < 0.001$), suggesting that researchers occupying brokerage positions within the co-authorship network tend to produce more scientific publications. In contrast, the quadratic term of betweenness centrality is negative and statistically significant ($\beta = -0.0478$, $p < 0.001$). The combination of a positive linear term and a negative quadratic term indicate a nonlinear relationship between brokerage positions and publication productivity, suggesting that the positive effect of betweenness centrality diminishes at higher levels of network centrality.

The remaining variables display patterns largely consistent with the baseline model results. Repeated collaboration (perc_repeat) shows a positive and highly significant association with publication output ($\beta = 0.1804$, $p < 0.001$), indicating that stable collaborative relationships are associated with higher levels of research productivity. Team embeddedness continues to exhibit a negative and statistically significant relationship with publication counts ($\beta = -0.1085$, $p = 0.007$), suggesting that highly cohesive collaboration structures may limit exposure to diverse knowledge sources. Similarly, the author–inventor indicator remains positive and statistically significant ($\beta = 0.0962$, $p = 0.001$), indicating that researchers engaged in both scientific publishing and patenting activities tend to demonstrate higher levels of publication productivity. In contrast, the industry collaboration indicator (bin_indust) does not appear to have a statistically significant effect in this specification.

Table 12 reports the results of an alternative specification including a quadratic term for group size. In this model, the linear term of group size is not statistically significant ($\beta = 0.0048$, $p = 0.819$), while the quadratic term is negative and statistically significant ($\beta = -0.0116$, $p = 0.015$). These results suggest the presence of a nonlinear relationship between team size and publication productivity. However, the overall improvement in model fit relative to the baseline specification appears limited.

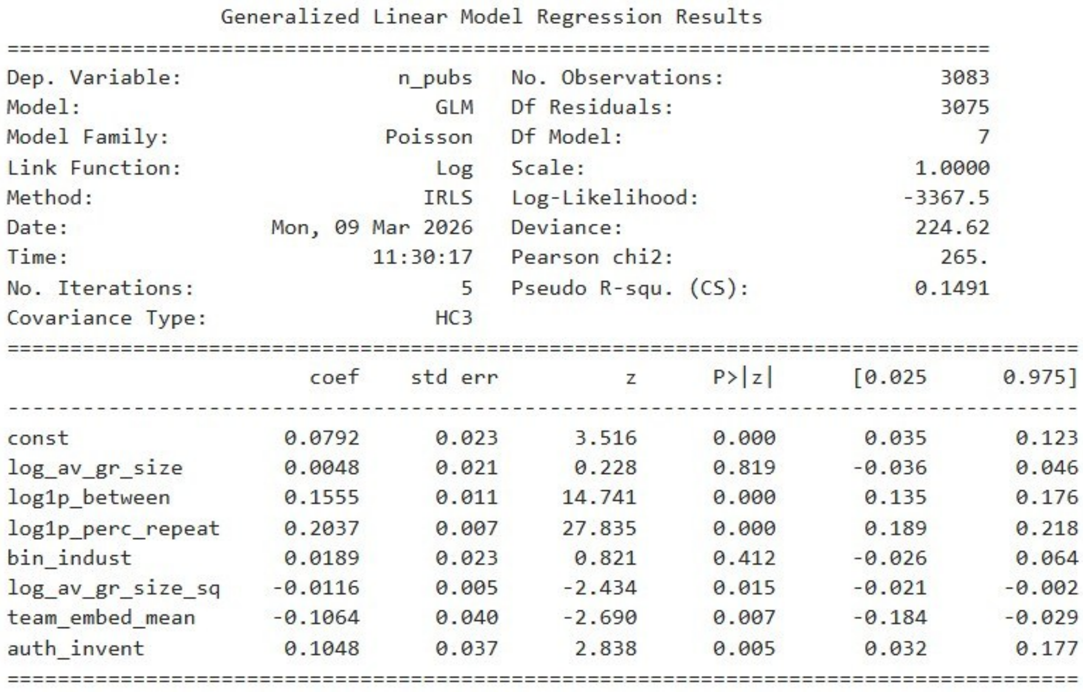
A comparison of model fit statistics indicates that the specification including the quadratic term for betweenness centrality provides a more noticeable improvement in model performance than the specification including the quadratic term for group size. In particular, the model incorporating the quadratic betweenness term shows a higher log-likelihood and a lower Pearson chi-square statistic compared with the baseline model. In contrast, the quadratic group size specification produces only minimal changes in these model fit indicators.

Overall, these results suggest that nonlinear effects are more pronounced for brokerage positions within the collaboration network than for team size, indicating that the relationship between network centrality and scientific productivity may follow a nonlinear pattern.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3075
Model Family:                 Poisson   Df Model:                            7
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3354.0
Date:                Mon, 09 Mar 2026   Deviance:                       197.58
Time:                        11:27:14   Pearson chi2:                     217.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1565
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1643      0.017      9.532      0.000       0.131       0.198
log_av_gr_size       -0.0706      0.009     -8.173      0.000      -0.087      -0.054
log1p_between         0.4025      0.037     10.749      0.000       0.329       0.476
log1p_perc_repeat     0.1804      0.007     26.370      0.000       0.167       0.194
bin_indust            0.0169      0.021      0.797      0.425      -0.025       0.058
log1p_between_sq     -0.0478      0.008     -6.268      0.000      -0.063      -0.033
team_embed_mean      -0.1085      0.040     -2.679      0.007      -0.188      -0.029
auth_invent           0.0962      0.030      3.178      0.001       0.037       0.156
=====================================================================================
```

Table 11. Quadratic betweenness model



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3075
Model Family:                 Poisson   Df Model:                            7
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3367.5
Date:                Mon, 09 Mar 2026   Deviance:                       224.62
Time:                        11:30:17   Pearson chi2:                     265.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1491
Covariance Type:                  HC3
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const                0.0792      0.023      3.516      0.000       0.035       0.123
log_av_gr_size       0.0048      0.021      0.228      0.819      -0.036       0.046
log1p_between        0.1555      0.011     14.741      0.000       0.135       0.176
log1p_perc_repeat    0.2037      0.007     27.835      0.000       0.189       0.218
bin_indust           0.0189      0.023      0.821      0.412      -0.026       0.064
log_av_gr_size_sq   -0.0116      0.005     -2.434      0.015      -0.021      -0.002
team_embed_mean     -0.1064      0.040     -2.690      0.007      -0.184      -0.029
auth_invent          0.1048      0.037      2.838      0.005       0.032       0.177
====================================================================================
```

Table 12. Quadratic team size model

To further investigate whether the effects of collaboration structures depend on researchers' engagement with industry and technological activities, additional interaction terms were introduced into the regression models. Prior research on scientific collaboration suggests that the impact of network position and collaboration structure may vary across different types of researchers and collaborative contexts (Fleming et al., 2007; Powell et al., 2005). Brokerage positions and industry engagement may jointly influence the ability of researchers to recombine diverse knowledge sources.

The first interaction term examined the joint effect of author–inventor status and betweenness centrality. Researchers who simultaneously engage in scientific publishing and patenting activities often operate at the interface between scientific and technological knowledge domains. To capture this mechanism, an interaction term between author–inventor status and betweenness centrality was included in the regression model.

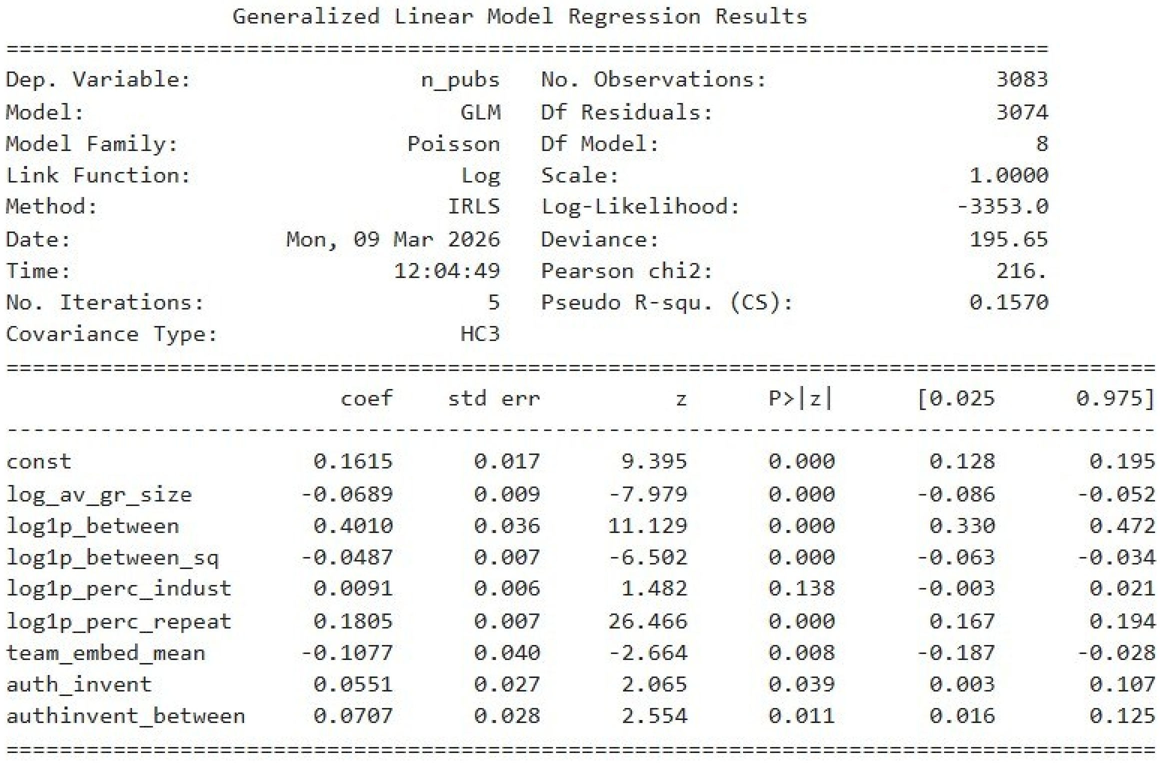
The estimation results in table 13 indicate that this interaction effect is positive and statistically significant (β = 0.0707, p = 0.011). This finding suggests that the positive relationship between brokerage positions and publication productivity is stronger for researchers who are also involved in patenting activities. In other words, author–inventor researchers appear to benefit more strongly from occupying brokerage positions within the co-authorship network. In addition to the statistical significance of the interaction term, the inclusion of this interaction also leads to a modest improvement in model fit. The log-likelihood increases from approximately −3354.0 in the baseline quadratic model to −3353.0, indicating a better fit of the model to the observed data. Correspondingly, the Akaike Information Criterion (AIC) decreases slightly, suggesting that the model including the interaction provides a better balance between explanatory power and model complexity.

The second interaction examined the relationship between group size and industry collaboration. Prior research on collaborative research teams suggests that while industry collaborations may provide access to additional resources and applied knowledge, larger research teams may also face increased coordination costs and communication challenges (Cummings & Kiesler, 2005; Wuchty et al., 2007). Consequently, the productivity benefits of industry collaboration may depend on the size of the research team. To examine this possibility, an interaction term between group size and the proportion of industry collaborators was introduced.

The results in table 14 show that this interaction effect is negative and statistically significant ($\beta = -0.0257$, $p < 0.001$). This finding indicates that the positive contribution of industry collaboration tends to diminish as research team size increases. One possible explanation is that collaborations involving both academic and industry partners may become more difficult to coordinate when the number of collaborators increases, which may reduce the efficiency of knowledge production. Similar to the previous interaction, the inclusion of the group size × industry interaction also improves model fit relative to the baseline model. The log-likelihood increases from approximately −3354.0 to about −3353.0, while the AIC decreases slightly, indicating an improved model specification.

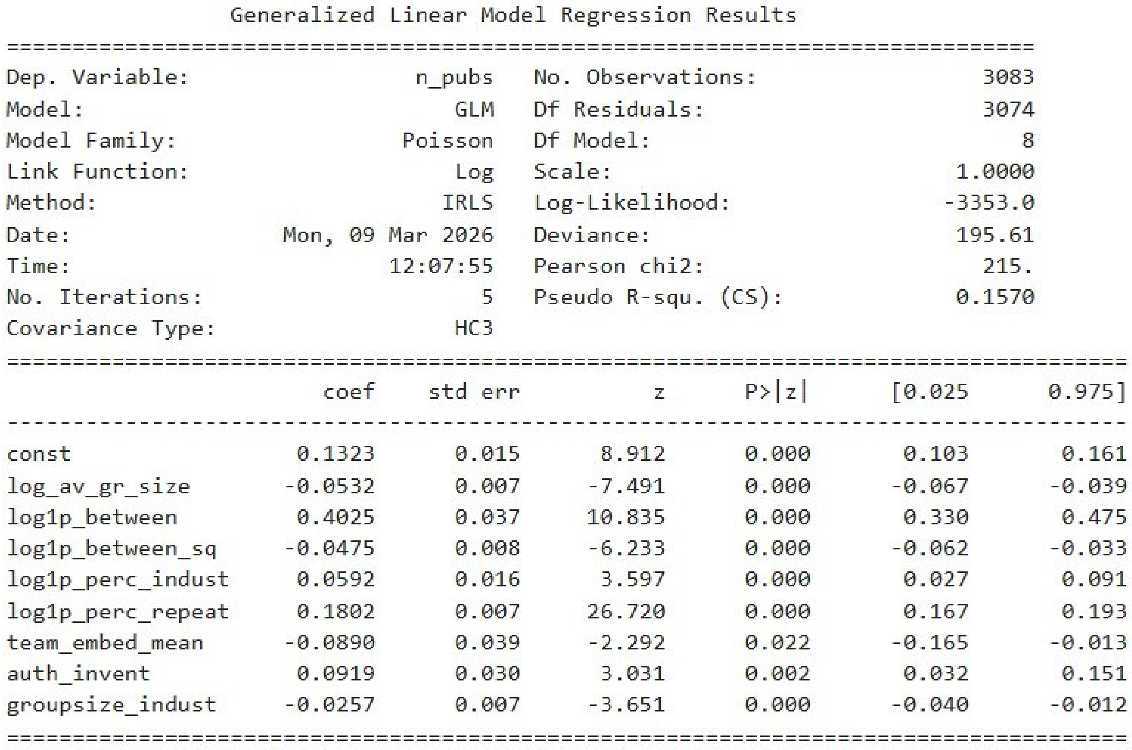
Finally, a model including both interaction terms simultaneously was estimated to assess whether these mechanisms operate jointly. The results show that both interaction effects remain statistically significant when included together in the model. Moreover, the model fit improves further, with the log-likelihood increasing to approximately −3352.3, which represents the highest log-likelihood value among the estimated models. The corresponding AIC also reaches its lowest level in this specification, indicating that the model including both interaction terms provides the best overall fit.

Taken together, these results suggest that the relationship between collaboration structures and scientific productivity is moderated by both researchers' technological engagement and the organizational composition of research teams. Brokerage positions appear to generate stronger productivity benefits for author–inventor researchers, while the positive effects of industry collaboration tend to weaken in very large research teams. Because both interaction terms remain statistically significant and lead to improvements in model fit, they are retained in the final model specification.



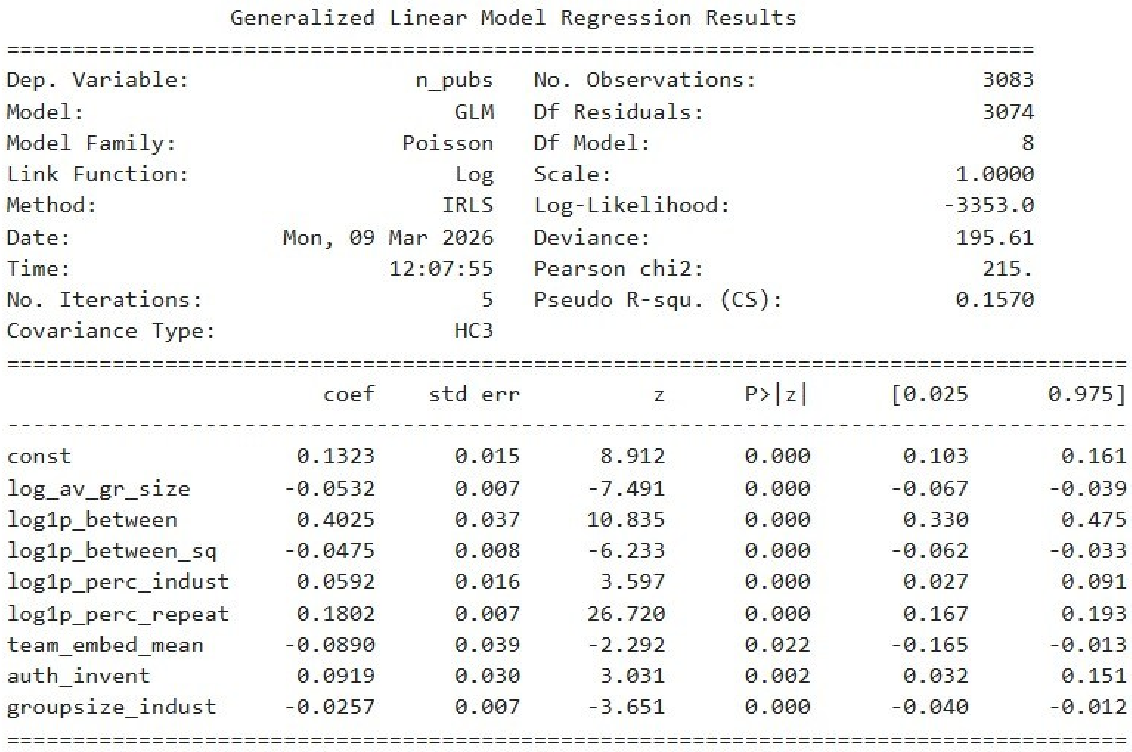
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3074
Model Family:                 Poisson   Df Model:                            8
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3353.0
Date:                Mon, 09 Mar 2026   Deviance:                       195.65
Time:                        12:04:49   Pearson chi2:                     216.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1570
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1615      0.017      9.395      0.000       0.128       0.195
log_av_gr_size       -0.0689      0.009     -7.979      0.000      -0.086      -0.052
log1p_between         0.4010      0.036     11.129      0.000       0.330       0.472
log1p_between_sq     -0.0487      0.007     -6.502      0.000      -0.063      -0.034
log1p_perc_indust     0.0091      0.006      1.482      0.138      -0.003       0.021
log1p_perc_repeat     0.1805      0.007     26.466      0.000       0.167       0.194
team_embed_mean      -0.1077      0.040     -2.664      0.008      -0.187      -0.028
auth_invent           0.0551      0.027      2.065      0.039       0.003       0.107
authinvent_between    0.0707      0.028      2.554      0.011       0.016       0.125
=====================================================================================
```

Table 13. Interaction of author-inventor and betweenness model



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3074
Model Family:                 Poisson   Df Model:                            8
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3353.0
Date:                Mon, 09 Mar 2026   Deviance:                       195.61
Time:                        12:07:55   Pearson chi2:                     215.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1570
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1323      0.015      8.912      0.000       0.103       0.161
log_av_gr_size       -0.0532      0.007     -7.491      0.000      -0.067      -0.039
log1p_between         0.4025      0.037     10.835      0.000       0.330       0.475
log1p_between_sq     -0.0475      0.008     -6.233      0.000      -0.062      -0.033
log1p_perc_indust     0.0592      0.016      3.597      0.000       0.027       0.091
log1p_perc_repeat     0.1802      0.007     26.720      0.000       0.167       0.193
team_embed_mean      -0.0890      0.039     -2.292      0.022      -0.165      -0.013
auth_invent           0.0919      0.030      3.031      0.002       0.032       0.151
groupsize_indust     -0.0257      0.007     -3.651      0.000      -0.040      -0.012
=====================================================================================
```

Table 14. Interaction of team size and percent industry model



```
                  Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 n_pubs   No. Observations:                 3083
Model:                            GLM   Df Residuals:                     3074
Model Family:                 Poisson   Df Model:                            8
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3353.0
Date:                Mon, 09 Mar 2026   Deviance:                       195.61
Time:                        12:07:55   Pearson chi2:                     215.
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1570
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1323      0.015      8.912      0.000       0.103       0.161
log_av_gr_size       -0.0532      0.007     -7.491      0.000      -0.067      -0.039
log1p_between         0.4025      0.037     10.835      0.000       0.330       0.475
log1p_between_sq     -0.0475      0.008     -6.233      0.000      -0.062      -0.033
log1p_perc_indust     0.0592      0.016      3.597      0.000       0.027       0.091
log1p_perc_repeat     0.1802      0.007     26.720      0.000       0.167       0.193
team_embed_mean      -0.0890      0.039     -2.292      0.022      -0.165      -0.013
auth_invent           0.0919      0.030      3.031      0.002       0.032       0.151
groupsize_indust     -0.0257      0.007     -3.651      0.000      -0.040      -0.012
=====================================================================================
```

Table 15. Final model

4.3.7 Discussion of Publication count Results

Table 15 presents the estimation results of the final Poisson regression model examining the relationship between collaboration structures, industry engagement, and scientific productivity, measured as the total number of publications produced by researchers. The model is estimated using a Poisson generalized linear model with a log-link function and robust standard errors.

The results indicate that betweenness centrality plays a positive and statistically significant role in explaining publication productivity. The coefficient of betweenness is positive and significant (β = 0.088, $p < 0.001$), suggesting that researchers occupying brokerage positions in the co-authorship network tend to produce a higher number of publications. Scientists located in brokerage positions often connect otherwise disconnected research groups, allowing them to access diverse knowledge flows and facilitate knowledge recombination. Such positions provide advantages in generating new collaboration opportunities and integrating heterogeneous knowledge, which can enhance scientific productivity (Burt, 2004; Fleming, Mingo & Chen, 2007). At the same time, the results reveal a nonlinear relationship between betweenness centrality and publication productivity. The quadratic term of betweenness is negative and statistically significant (β = -0.004, $p < 0.05$), indicating diminishing returns associated with extremely high brokerage positions. While moderate levels of brokerage may facilitate knowledge recombination and collaboration opportunities, excessively central brokerage positions may increase coordination burdens and information overload, potentially limiting researchers' ability to convert network advantages into additional publications (Uzzi & Spiro, 2005; Fleming et al., 2007).

The analysis also shows a negative and statistically significant effect of group size on publication output (β = -0.094, $p < 0.001$). This finding suggests that larger research teams may be associated with lower individual publication productivity. Although larger teams provide access to complementary expertise and research resources, they may also generate coordination costs and communication complexity that reduce the marginal productivity of individual researchers (Cummings & Kiesler, 2005; Jones, Wuchty & Uzzi, 2008).

In contrast, repeated collaboration exhibits a strong positive relationship with publication output (β = 0.454, $p < 0.001$), indicating that stable collaborative relationships improve research productivity by facilitating trust, coordination, and efficient knowledge exchange among collaborators (Gulati, 1995; Uzzi, 1997).

The results further show that team embeddedness has a negative and statistically significant association with publication productivity (β = -0.262, $p < 0.001$). Highly cohesive collaboration structures may limit exposure to diverse knowledge sources, as dense networks often generate redundant information flows and reduce opportunities for novel knowledge recombination (Burt, 2004; Uzzi, 1997).

From the perspective of industry engagement, the baseline results indicate that the direct effect of the proportion of industry collaborators (perc_indust) is not statistically significant (β = -0.058, $p > 0.1$). This suggests that the mere presence of industry collaborators within research teams does not automatically translate into higher academic publication productivity. Previous studies have shown that university–industry collaboration often generates multiple forms of output, including technological innovation and patenting activity, which may not necessarily increase academic publication counts (Perkmann et al., 2013; Breschi & Catalini, 2010). The results also indicate that author–inventor status has a positive and statistically significant direct effect on publication productivity. Researchers who are simultaneously involved in both scientific publishing and patenting activities tend to produce a higher number of academic publications compared with those who are not engaged in patenting. This finding highlights the role of hybrid researchers who operate at the interface between academic science and technological innovation. Author–inventors often participate in both scientific and technological knowledge networks, which may provide access to a broader range of ideas, resources, and collaboration opportunities. Through these connections, they may benefit from knowledge spillovers between academic research and technological development, which can enhance scientific productivity (Fleming & Sorenson, 2004; Breschi & Catalini, 2010; Perkmann et al., 2013).

However, the interaction analysis reveals that the impact of industry collaboration depends strongly on the structure of research teams. The interaction between group size and the proportion of industry collaborators is negative and statistically significant ($\beta = -0.107$, $p < 0.01$). This result suggests that the productivity implications of industry collaboration vary depending on team size. In larger research teams, the presence of industry collaborators may increase coordination complexity and introduce differences between academic and commercial research objectives, which may reduce publication productivity (Powell, Koput & Smith-Doerr, 1996; Perkmann et al., 2013).

Furthermore, the results indicate that author–inventor status interacts positively with betweenness centrality ($\beta = 0.054$, $p < 0.05$). This finding suggests that researchers who are simultaneously involved in scientific publishing and patenting activities benefit more strongly from brokerage positions in collaboration networks. Author–inventors often operate at the intersection of scientific and technological knowledge domains, enabling them to connect heterogeneous knowledge communities and facilitate knowledge transfer between academia and industry (Fleming & Sorenson, 2004; Breschi & Catalini, 2010).

Overall, the findings highlight that the relationship between industry engagement and scientific productivity is not purely direct but depends on collaboration structures within research networks. While industry collaboration alone does not significantly increase publication output, its effects become visible when combined with specific network positions and team structures. In particular, the results suggest that the benefits of industry engagement are mediated by brokerage positions and moderated by the size of research teams. These findings underline the importance of considering both network structure and forms of industry interaction simultaneously when analyzing the impact of university–industry collaboration on scientific productivity.

4.4 Citation Mean

In addition to scientific productivity measured by publication counts, this study also examines scientific impact using citation-based indicators. Citations are widely used in scientometric and bibliometric research as a proxy for the influence and visibility of scientific work within the academic community. Because citations reflect how often research is referenced by subsequent studies, they are commonly interpreted as an indicator of the recognition and impact of scientific contributions.

Citation-based indicators are extensively used to evaluate research impact at the level of publications, authors, institutions, and research fields. For example, Zhang and Wu (2024) analyze citation impact across research areas and highlight the central role of citation counts in measuring scientific influence. Similarly, Ke et al. (2023) employ citation-based indicators in a network-based framework to assess research impact across disciplines. In addition, Chen (2023) examines citation metrics within large citation networks and confirms their importance for evaluating scientific influence.

Consistent with this literature, the present study uses average citations per author as an indicator of scientific impact. While publication counts reflect scientific productivity, citation indicators capture the extent to which researchers' work influences subsequent research. By analyzing citation performance alongside publication productivity, this study investigates whether researchers' network position and collaboration with industry partners are associated not only with higher research output but also with greater scientific impact.

4.4.1 Sample Preparation and Data Screening

The general procedures for dataset construction, variable definition, and initial data screening were described in the previous part and are not repeated here. However, additional data preparation steps were required for the citation-based analysis due to the presence of missing citation values.

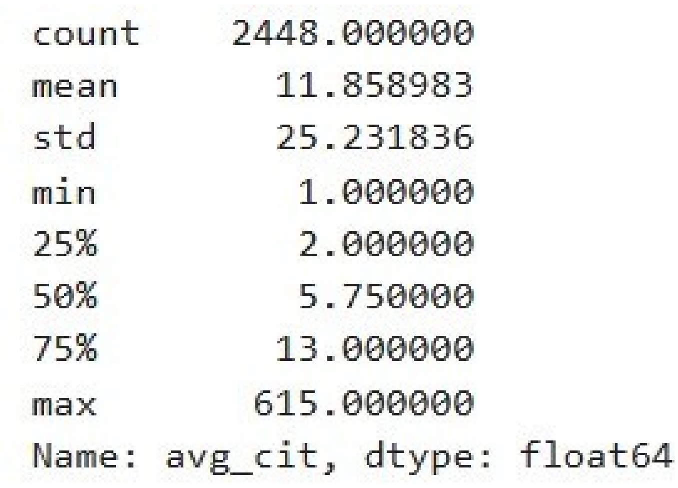
In the original dataset, the analytical sample consisted of 3,083 unique Canadian AI authors identified from Scopus publications. However, approximately 20% of the observations contained missing values for citation counts. Because citation impact is the dependent variable in this part of the analysis, observations with missing citation values cannot provide valid information for model estimation. For this reason, these observations were removed from the citation analysis, resulting in a final analytical sample of 2,448 authors.

Importantly, missing citation values were not replaced with zeros. Treating missing citations as zero would incorrectly imply that the corresponding publications received no citations, whereas the citation information was simply unavailable in the dataset. Replacing missing values with zero would therefore introduce measurement error and bias the estimation of citation impact. Consequently, observations with missing citation values were excluded from the analysis to preserve the validity of the dependent variable.

Because the removal of these observations changes the composition of the analytical dataset, the distributional characteristics of the variables were re-examined for the reduced sample. In particular, the distributions of all independent variables were reassessed to verify whether the transformations applied in the publication analysis remained appropriate. The inspection of these variables confirmed that the previously applied transformations—such as the logarithmic transformations of degree centrality, betweenness centrality, repeat collaboration intensity, and average group size—remained suitable for the citation analysis.

The dependent variable used in this section is the average number of citations per author (avg_cit). As it is seen in table 16, the distribution of this variable exhibits substantial right-skewness and a long upper tail, with a small number of authors receiving very high citation values compared with many researchers. Descriptive statistics confirm this pattern: while the median citation level remains relatively low, the maximum value is considerably larger, indicating a highly skewed distribution.

To reduce skewness and stabilize variance in the regression analysis, the citation variable was transformed using the natural logarithm of average citations. This transformation allows the regression models to better approximate linear relationships while maintaining interpretability of the estimated coefficients.



```
count    2448.000000
mean       11.858983
std        25.231836
min         1.000000
25%         2.000000
50%         5.750000
75%        13.000000
max       615.000000
Name: avg_cit, dtype: float64
```

Table 16. Descriptive statistics for the average citation per author

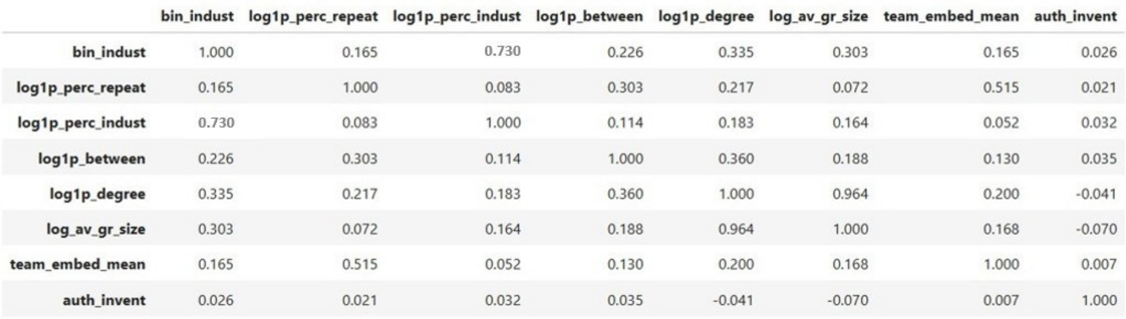
4.4.2 Correlation

Because the citation analysis is conducted on a reduced sample after removing observations with missing citation values, the correlations were recalculated for the filtered dataset. The exclusion of these observations changed the composition of the analytical sample and therefore required re-examining the relationships among the variables used in the regression analysis.

In particular, the correlation analysis was conducted to identify potential multicollinearity among the explanatory variables prior to estimating the OLS regression models. Addressing multicollinearity is important in OLS estimation because highly correlated independent variables can inflate standard errors, reduce the statistical reliability of coefficient estimates, and make the interpretation of individual effects difficult.

Consistent with the findings reported in the publication analysis, the correlation results in table 17 indicate strong associations between degree centrality and average group size, as well as between binary industry collaboration and the percentage of industry collaborators. Because these pairs of variables capture closely related aspects of network position and industry engagement, including them simultaneously in the same regression specification could introduce multicollinearity problems.

Therefore, following the same modeling strategy used in the publication analysis, these highly correlated variables were included in separate model specifications, allowing the regression models to isolate their effects while avoiding multicollinearity issues.



	bin_indust	log1p_perc_repeat	log1p_perc_indust	log1p_between	log1p_degree	log_av_gr_size	team_embed_mean	auth_invent
bin_indust	1.000	0.165	0.730	0.226	0.335	0.303	0.165	0.026
log1p_perc_repeat	0.165	1.000	0.083	0.303	0.217	0.072	0.515	0.021
log1p_perc_indust	0.730	0.083	1.000	0.114	0.183	0.164	0.052	0.032
log1p_between	0.226	0.303	0.114	1.000	0.360	0.188	0.130	0.035
log1p_degree	0.335	0.217	0.183	0.360	1.000	0.964	0.200	-0.041
log_av_gr_size	0.303	0.072	0.164	0.188	0.964	1.000	0.168	-0.070
team_embed_mean	0.165	0.515	0.052	0.130	0.200	0.168	1.000	0.007
auth_invent	0.026	0.021	0.032	0.035	-0.041	-0.070	0.007	1.000

Table 17. Pearson Correlation Matrix of Independent Variables

4.4.3 Model Selection

The choice of the regression model depends on the nature and distribution of the dependent variable. In this study, the dependent variable is average citations per author, which represents a continuous indicator of scientific impact.

Descriptive statistics show that the distribution of average citations is strongly right skewed, with a small number of authors receiving very high citation values compared with many researchers. Such skewness is common in citation data and may affect the stability and interpretation of regression estimates.

To address this issue, the dependent variable was transformed using the natural logarithm of average citations. This transformation reduces skewness and stabilizes variance, allowing the regression model to better approximate linear relationships between the dependent and explanatory variables.

Based on these characteristics, the citation models were estimated using Ordinary Least Squares (OLS) regression applied to the log-transformed average citation variable. Robust standard errors (HC3) were used to account for potential heteroskedasticity commonly observed in citation data.

4.4.4 Theoretical Model Specification

The theoretical framework of this study examines whether researchers’ collaboration network characteristics and industry engagement may be associated with their scientific impact, measured by average citations per author. Prior research in scientometrics suggests that citation performance may be influenced not only by research productivity but also by the structure of collaboration networks and cross-sector collaboration patterns. However, the direction and magnitude of these relationships are not always straightforward and may vary depending on the characteristics of collaboration networks.

One potential factor that may influence citation impact is the position of researchers within collaboration networks. Scholars occupying more central positions in co-authorship networks may gain greater visibility, broader access to information, and more collaboration opportunities. These advantages could potentially increase the diffusion of their research outputs and therefore lead to higher citation impact (Abbasi et al., 2012; Yan & Ding, 2009). At the same time, however, highly central researchers may also face coordination costs associated with managing large collaboration networks, which could limit the marginal benefits of additional collaborations. For this reason, the relationship between degree centrality and citation impact may not necessarily be strictly positive and may vary depending on the structure of collaboration networks.

In addition to degree centrality, the model considers betweenness centrality, which reflects the extent to which a researcher connects otherwise separated groups of collaborators. Researchers occupying such bridging positions may facilitate knowledge flows between different scientific communities and potentially increase the visibility of their research outputs (Freeman, 1979; Newman, 2001). However, acting as a bridge between multiple groups may also involve higher coordination costs or weaker ties with collaborators, which could affect research productivity or visibility. Consequently, the overall relationship between betweenness centrality and citation impact may be positive in some contexts but less clear in others.

The stability of collaboration relationships may represent another factor influencing scientific impact. Repeated collaboration may strengthen trust among collaborators and facilitate knowledge integration, potentially leading to more visible or influential research outputs (Uzzi et al., 2013). At the same time, repeatedly collaborating with the same partners could also reduce exposure to new ideas and limit the diversity of knowledge sources. As a result, the relationship between repeated collaboration and citation impact may vary depending on the balance between stability and diversity within research networks.

Another important dimension of collaboration networks concerns the interaction between academia and industry. Collaboration with industry partners may provide researchers with access to additional resources, applied knowledge, and new research questions that could enhance the relevance and visibility of scientific outputs (Perkmann et al., 2013).

To capture different aspects of industry engagement, the model includes two indicators of industry collaboration. The first is the percentage of industry collaborators (perc_indust), which reflects the extent to which industry actors are present within a researcher's collaboration network. A higher proportion of industry collaborators may increase exposure to applied knowledge and practical research problems, which could enhance the visibility and societal relevance of research outputs and potentially influence citation impact (Perkmann et al., 2013; Abramo et al., 2011). At the same time, the relationship may not be strictly positive, as research conducted with strong industry orientation may sometimes focus on more applied outcomes that receive different levels of attention within academic citation networks.

The model also includes a binary indicator of industry collaboration (bin_indust) that captures whether a researcher collaborates with industry partners. This variable allows the analysis to examine whether simply engaging in industry collaboration is associated with differences in citation performance compared with researchers who collaborate exclusively within academia.

In addition, the analysis considers the presence of author–inventors, defined as researchers who are active both in scientific publishing and in patenting activities. Author–inventors represent a particularly strong form of interaction between scientific research and technological development. Because these researchers operate at the interface between academic and technological knowledge systems, they may benefit from broader knowledge flows and increased visibility across different innovation networks (Perkmann et al., 2013; Meyer, 2006). As a result, their scientific publications may receive higher attention and citation impact in some contexts.

By incorporating these indicators, the model allows the analysis to examine whether different forms of industry engagement—participation in industry collaboration, the intensity of such collaboration, and involvement in patenting activities—may be associated with variations in citation impact.

In addition to examining direct relationships, the model also considers the possibility that the association between certain network characteristics and citation impact may be nonlinear. While increasing collaboration or occupying more central network positions may improve access to knowledge and research visibility, these benefits may not increase indefinitely. At higher levels of network connectivity, coordination costs and the complexity of managing multiple collaborations may reduce the marginal advantages of additional ties.

For this reason, the analysis explores potential nonlinear relationships for key network variables, particularly degree centrality and betweenness centrality. Degree centrality reflects the number of collaborative ties maintained by a researcher. Although greater collaboration may initially increase research visibility and knowledge access, very large collaboration networks may involve higher coordination costs, which could reduce the marginal benefits of additional collaborations (Abbasi et al., 2012). To examine this possibility, a quadratic term for degree centrality is included in the model. Similarly, betweenness centrality captures the extent to which a researcher connects otherwise separated groups within the collaboration network. Such bridging positions may facilitate knowledge diffusion and increase research visibility, but managing highly complex bridging roles may also influence research outcomes in different ways (Freeman, 1979; Newman, 2001). Therefore, quadratic terms are also tested for betweenness centrality to explore whether its relationship with citation impact varies across different levels of network embeddedness.

In addition to examining direct and nonlinear relationships, the model also considers the possibility that the influence of network position on citation impact may vary depending on researchers' industry collaboration status. Prior research suggests that collaboration across institutional boundaries can influence the diffusion and visibility of scientific knowledge (Perkmann et al., 2013; Abramo et al., 2011). As a result, the relationship between network position and scientific impact may differ between researchers who collaborate with industry partners and those who operate solely within academic networks.

Researchers who collaborate with industry may gain access to additional knowledge sources, practical research problems, and broader innovation networks. These connections could potentially strengthen the benefits of occupying central positions in collaboration networks by expanding channels for knowledge diffusion and increasing the visibility of research outputs (Perkmann et al., 2013; Meyer, 2006). However, the extent to which industry collaboration modifies the relationship between network position and citation impact remains an empirical question.

To explore this possibility, the analysis includes interaction terms between industry collaboration indicators and network centrality measures, particularly between industry collaboration and degree centrality. Including these interaction terms allows the model to examine whether the association between network position and citation impact differs between researchers engaged in industry collaboration and those whose collaborations occur exclusively within academic networks.

4.4.5 Baseline Models Specifications

In this part of the analysis, the dependent variable is scientific impact, measured by the average number of citations per author. Because the distribution of average citations is strongly right skewed, the dependent variable was transformed using the natural logarithm of average citations to reduce skewness and improve the linear specification of the model. Thus, the baseline models examine the relationship between the explanatory variables and log-transformed average citation.

The relationship is estimated using Ordinary Least Squares (OLS) regression, where the log-transformed citation indicator is modeled as a function of researchers' network characteristics, industry collaboration variables, and control variables.

The baseline specification can be written as follows:

$$\ln(\mathrm{AvgCitation}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

where:

$\ln(AvgCitation_i)$ is the natural logarithm of the average citation of researcher i,

X_{1i} , ….,X_{ki} represent the explanatory variables,

B_0 is the intercept,

B_2, …., B_k are the coefficients to be estimated,

and ε_i is the error term.

Following the correlation analysis, some explanatory variables were found to be highly correlated. Strong correlations were observed between degree centrality and average group size, as well as between binary industry collaboration (bin_indust) and the percentage of industry collaborators (perc_indust). Including these variables simultaneously in the same regression model could lead to multicollinearity problems and reduce the reliability of coefficient estimates.

To address this issue, and like the publication analysis, the citation models are estimated using four separate baseline specifications. In these models, degree centrality and average group size are included in alternative specifications, and binary industry collaboration and the percentage of industry collaborators are also introduced in separate models. This approach allows the analysis to examine the effects of network position and industry collaboration on citation impact while avoiding multicollinearity among highly correlated variables.

4.4.6 Baseline Models Estimation and Comparison

As reported in Tables 18–21, which present the results of the four baseline regression models, several explanatory variables show consistent relationships with the dependent variable, average citation impact. Both degree centrality and average group size exhibit positive and statistically significant associations with citation performance across the baseline specifications. Similar positive relationships between collaboration intensity or research team size and citation impact have been documented in previous studies (Wuchty et al., 2007; Abbasi et al., 2012).

The regression results also indicate a positive and statistically significant relationship between betweenness centrality and citation impact across the models reported in all four tables. Comparable positive associations between bridging positions in collaboration networks and scientific impact indicators have been identified in earlier research (Freeman, 1979; Newman, 2001).

Similarly, the variables capturing industry collaboration show positive relationships with citation impact in the baseline models. Both the binary indicator of industry collaboration (bin_indust) and the percentage of industry collaborators (perc_indust) display positive coefficients in the regression results reported in Tables 18–21. Previous studies have also reported positive relationships between university–industry collaboration and citation impact (Perkmann et al., 2013; Abramo et al., 2011).

In addition, the coefficient associated with author–inventor status is positive across the baseline specifications reported in all four tables and approaches statistical significance. Similar positive relationships between patenting activity and scientific impact have been documented in earlier studies (Meyer, 2006).

By contrast, the variables representing repeat collaboration intensity and team embeddedness do not show statistically significant relationships with citation impact in the baseline models. Similar mixed or insignificant effects for repeated collaboration have also been reported in the scientometric literature (Uzzi et al., 2013).

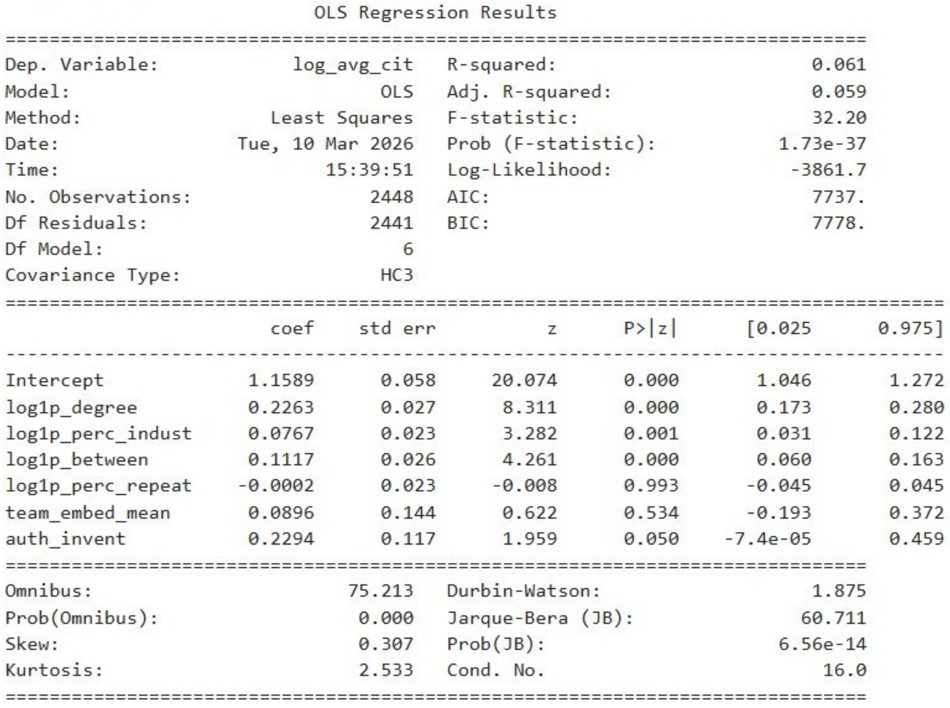
To identify the most appropriate specification among the four baseline models, several model selection criteria commonly used in OLS regression were considered, including Adjusted R^2, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). Adjusted R^2 reflects the explanatory power of the model while accounting for the number of explanatory variables, whereas AIC and BIC evaluate the trade-off between model fit and model complexity, with lower values indicating a preferable specification (Burnham & Anderson, 2002).

The comparative results for the four baseline models are summarized in Table 22. Among the candidate specifications, Model M2_degree_bin demonstrates the best overall performance. This model reports the highest Adjusted R^2 (0.061483) and the lowest AIC (7731.29) and BIC (7771.92) values among the four models. The remaining specifications show slightly weaker performance. Model M4_grsize_bin yields an Adjusted R^2 of 0.059182, with AIC = 7737.29 and BIC = 7777.91, while M1_degree_perc reports an Adjusted R^2 of 0.059117, with AIC = 7737.46 and BIC = 7778.08. Finally, M3_grsize_perc shows the lowest explanatory power among the four models, with an Adjusted R^2 of 0.056667 and the highest information criteria values (AIC = 7743.82, BIC = 7784.45).

Although the differences in Adjusted R^2 across the models are relatively small, the information criteria (AIC and BIC) provide a clearer basis for model comparison because they simultaneously consider model fit and model parsimony (Burnham & Anderson, 2002). Based on these criteria, Model M2_degree_bin provides the most efficient balance between explanatory power and model simplicity and is therefore selected as the preferred baseline specification.

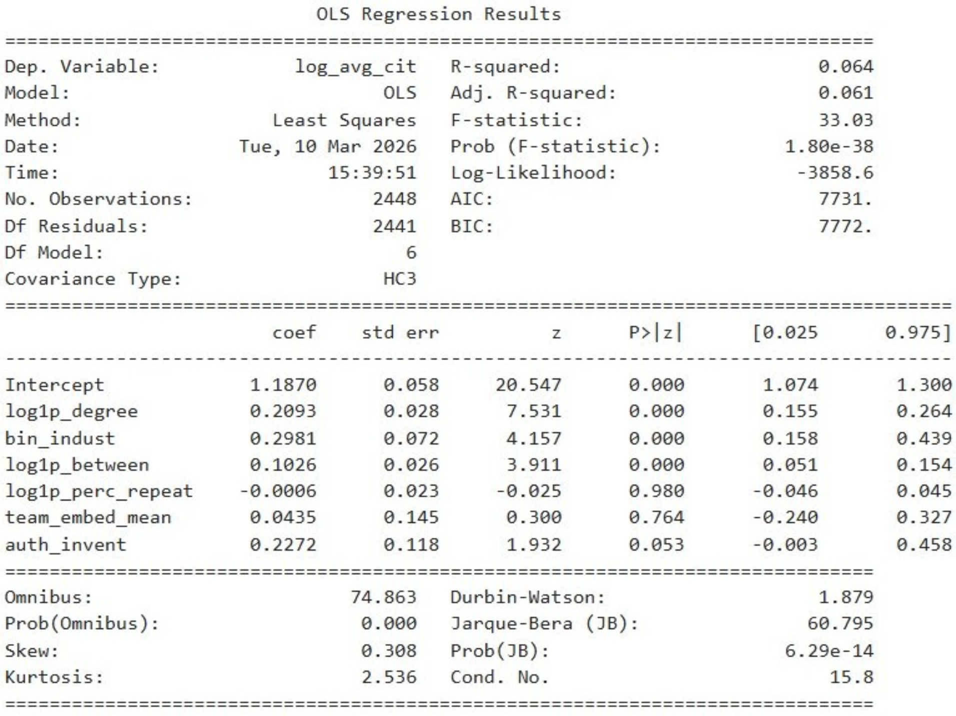
It is also important to note that relatively low R² values are common in citation-based studies, where research impact is influenced by many factors that are difficult to fully capture in empirical models. Previous studies analyzing citation performance frequently report modest explanatory power even when statistically significant relationships are identified (Bornmann & Daniel, 2008; Abbasi et al., 2012). Consequently, the Adjusted R² values observed in this study are consistent with the broader scientometric literature and do not indicate a weakness of the model specification.

Accordingly, Model M2_degree_bin is selected as the reference model for the subsequent analysis of nonlinear and interaction effects.



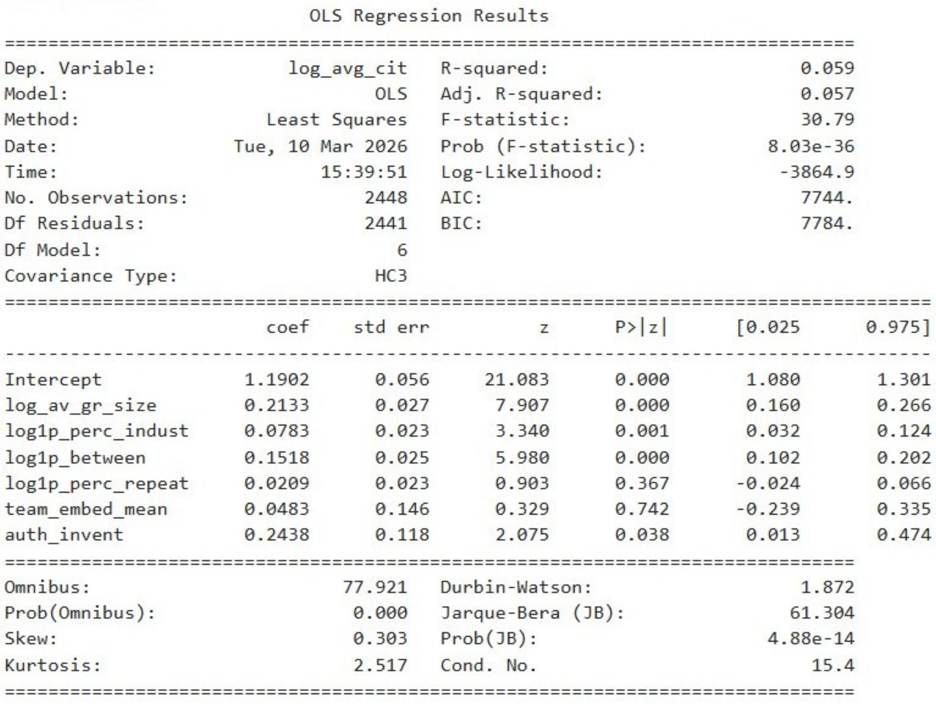
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.061
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     32.20
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           1.73e-37
Time:                        15:39:51   Log-Likelihood:                -3861.7
No. Observations:                2448   AIC:                             7737.
Df Residuals:                    2441   BIC:                             7778.
Df Model:                           6
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             1.1589      0.058     20.074      0.000       1.046       1.272
log1p_degree          0.2263      0.027      8.311      0.000       0.173       0.280
log1p_perc_indust     0.0767      0.023      3.282      0.001       0.031       0.122
log1p_between         0.1117      0.026      4.261      0.000       0.060       0.163
log1p_perc_repeat    -0.0002      0.023     -0.008      0.993      -0.045       0.045
team_embed_mean       0.0896      0.144      0.622      0.534      -0.193       0.372
auth_invent           0.2294      0.117      1.959      0.050    -7.4e-05       0.459
==============================================================================
Omnibus:                       75.213   Durbin-Watson:                   1.875
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               60.711
Skew:                           0.307   Prob(JB):                     6.56e-14
Kurtosis:                       2.533   Cond. No.                         16.0
==============================================================================
```

Table18. Baseline OLS regression results- M1



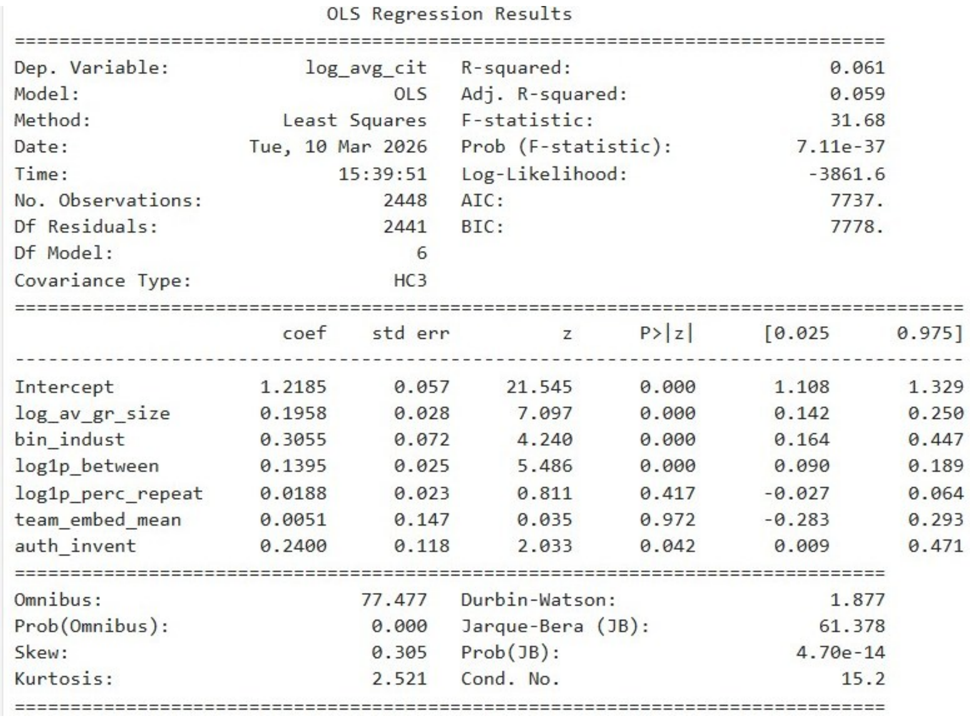
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.064
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     33.03
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           1.80e-38
Time:                        15:39:51   Log-Likelihood:                -3858.6
No. Observations:                2448   AIC:                             7731.
Df Residuals:                    2441   BIC:                             7772.
Df Model:                           6
Covariance Type:                  HC3
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              1.1870      0.058     20.547      0.000       1.074       1.300
log1p_degree           0.2093      0.028      7.531      0.000       0.155       0.264
bin_indust             0.2981      0.072      4.157      0.000       0.158       0.439
log1p_between          0.1026      0.026      3.911      0.000       0.051       0.154
log1p_perc_repeat     -0.0006      0.023     -0.025      0.980      -0.046       0.045
team_embed_mean        0.0435      0.145      0.300      0.764      -0.240       0.327
auth_invent            0.2272      0.118      1.932      0.053      -0.003       0.458
==============================================================================
Omnibus:                       74.863   Durbin-Watson:                   1.879
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               60.795
Skew:                           0.308   Prob(JB):                     6.29e-14
Kurtosis:                       2.536   Cond. No.                         15.8
==============================================================================
```

Table 19. Baseline OLS regression results- M2



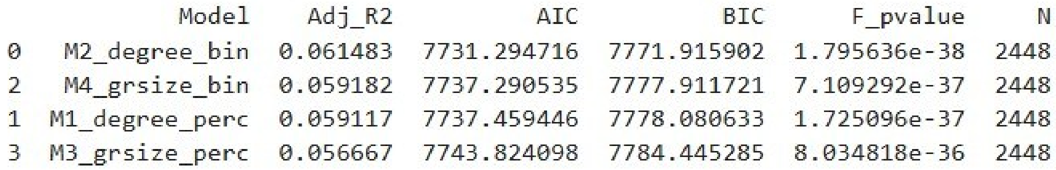
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.059
Model:                            OLS   Adj. R-squared:                  0.057
Method:                 Least Squares   F-statistic:                     30.79
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           8.03e-36
Time:                        15:39:51   Log-Likelihood:                -3864.9
No. Observations:                2448   AIC:                             7744.
Df Residuals:                    2441   BIC:                             7784.
Df Model:                           6
Covariance Type:                  HC3
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              1.1902      0.056     21.083      0.000       1.080       1.301
log_av_gr_size         0.2133      0.027      7.907      0.000       0.160       0.266
log1p_perc_indust      0.0783      0.023      3.340      0.001       0.032       0.124
log1p_between          0.1518      0.025      5.980      0.000       0.102       0.202
log1p_perc_repeat      0.0209      0.023      0.903      0.367      -0.024       0.066
team_embed_mean        0.0483      0.146      0.329      0.742      -0.239       0.335
auth_invent            0.2438      0.118      2.075      0.038       0.013       0.474
==============================================================================
Omnibus:                       77.921   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.304
Skew:                           0.303   Prob(JB):                     4.88e-14
Kurtosis:                       2.517   Cond. No.                         15.4
==============================================================================
```

Table 20. Baseline OLS regression results- M3



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.061
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     31.68
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           7.11e-37
Time:                        15:39:51   Log-Likelihood:                -3861.6
No. Observations:                2448   AIC:                             7737.
Df Residuals:                    2441   BIC:                             7778.
Df Model:                           6
Covariance Type:                  HC3
====================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept           1.2185      0.057     21.545      0.000       1.108       1.329
log_av_gr_size      0.1958      0.028      7.097      0.000       0.142       0.250
bin_indust          0.3055      0.072      4.240      0.000       0.164       0.447
log1p_between       0.1395      0.025      5.486      0.000       0.090       0.189
log1p_perc_repeat   0.0188      0.023      0.811      0.417      -0.027       0.064
team_embed_mean     0.0051      0.147      0.035      0.972      -0.283       0.293
auth_invent         0.2400      0.118      2.033      0.042       0.009       0.471
==============================================================================
Omnibus:                       77.477   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.378
Skew:                           0.305   Prob(JB):                     4.70e-14
Kurtosis:                       2.521   Cond. No.                         15.2
==============================================================================
```

Table 21. Baseline OLS regression results- M4



```
             Model    Adj_R2          AIC          BIC      F_pvalue     N
0    M2_degree_bin  0.061483  7731.294716  7771.915902  1.795636e-38  2448
2    M4_grsize_bin  0.059182  7737.290535  7777.911721  7.109292e-37  2448
1  M1_degree_perc  0.059117  7737.459446  7778.080633  1.725096e-37  2448
3  M3_grsize_perc  0.056667  7743.824098  7784.445285  8.034818e-36  2448
```

Table 22. Baseline models comparison

4.4.7 Extended Model Specifications

While the baseline models provide an initial understanding of the relationship between collaboration network characteristics and citation impact, they assume that the effects of explanatory variables are linear and constant across all observations. However, in studies of scientific collaboration networks, several scholars have suggested that the influence of network position on research performance may be nonlinear or may vary depending on other collaboration characteristics (Abbasi et al., 2012; Uzzi et al., 2013). Similarly, interaction effects may exist when the influence of one explanatory variable depends on the level of another variable. Considering such nonlinearities and interaction mechanisms can therefore provide a more accurate representation of collaboration dynamics and their relationship with scientific impact.

From a theoretical perspective, nonlinear effects are particularly plausible for network centrality measures, since the benefits of collaboration networks may not increase proportionally as network connectivity grows. In this study, two central network indicators—degree centrality and betweenness centrality—were selected for nonlinear examination. Degree centrality captures the extent of an author's direct collaborative ties, whereas betweenness centrality reflects the extent to which an author occupies bridging positions within the collaboration network. Previous studies have suggested that both types of centralities may exhibit nonlinear relationships with research performance due to mechanisms such as increasing coordination costs or diminishing marginal returns from additional collaborations (Freeman, 1979; Abbasi et al., 2012). To examine potential nonlinear effects, quadratic terms were estimated for both degree centrality and betweenness centrality. As reported in table 23, the coefficient associated with the quadratic term of betweenness centrality is statistically insignificant, indicating that no meaningful nonlinear relationship can be identified between betweenness centrality and citation impact in the present dataset. In contrast, the quadratic term of degree centrality is statistically significant and negative as it is shown in table 24, suggesting the presence of a nonlinear relationship between degree centrality and citation impact. Based on these results, only the quadratic specification for degree centrality is retained in the extended model.

More specifically, the positive linear coefficient of degree centrality combined with the negative coefficient of its quadratic term indicates an inverted-U shaped relationship between degree centrality and citation impact. This pattern suggests that citation impact increases with collaboration intensity up to a certain point, after which the marginal benefit of additional collaborations declines. Similar inverted-U patterns between collaboration intensity and scientific performance have been reported in previous studies of collaboration networks (Abbasi et al., 2012; Uzzi et al., 2013).

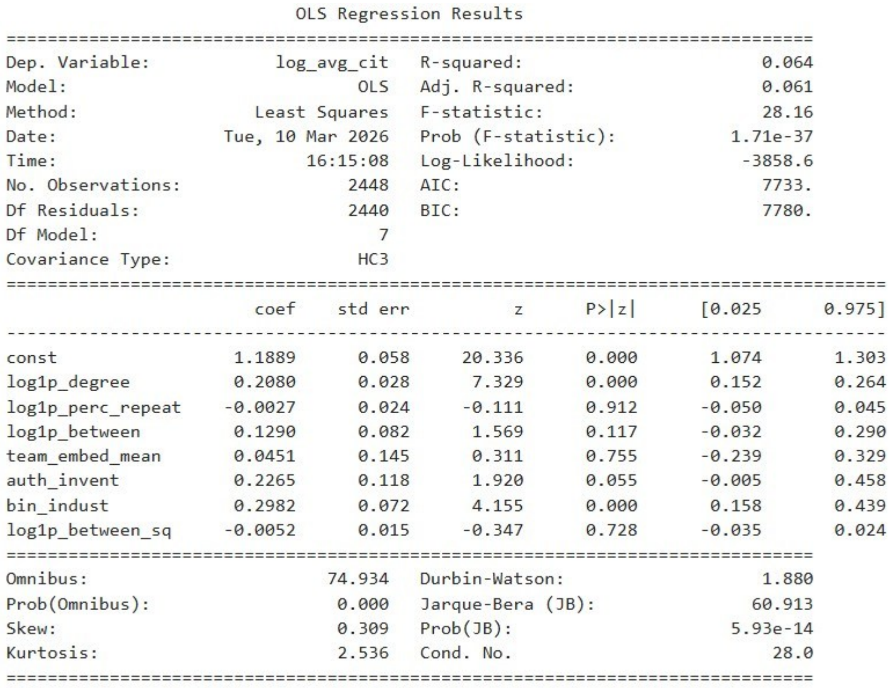
In addition to nonlinear effects, the extended model also examines whether the relationship between the explanatory variables and citation impact may vary depending on scientists' engagement with industry. Previous research suggests that collaboration with industry can benefit from centrality in terms of research outcomes (Perkmann et al., 2013; Abramo et al., 2011). Consequently, the effect of researchers who engage in industry collaboration on citation may differ based on their network positions.

Based on this reasoning, the analysis tests interaction terms between degree centrality/betweenness centrality and industry-related variables to examine better the influence of industry-related researchers on citation.

The regression results reported in Tables 25–28 show that the interaction terms involving author–inventor status are statistically insignificant. In particular, the coefficients associated with the interaction between author inventor and degree centrality as well as the interaction between author inventor and betweenness centrality do not reach statistical significance. These results suggest that the relationship between the status of author inventor and citation impact does not substantially differ for various centralities.

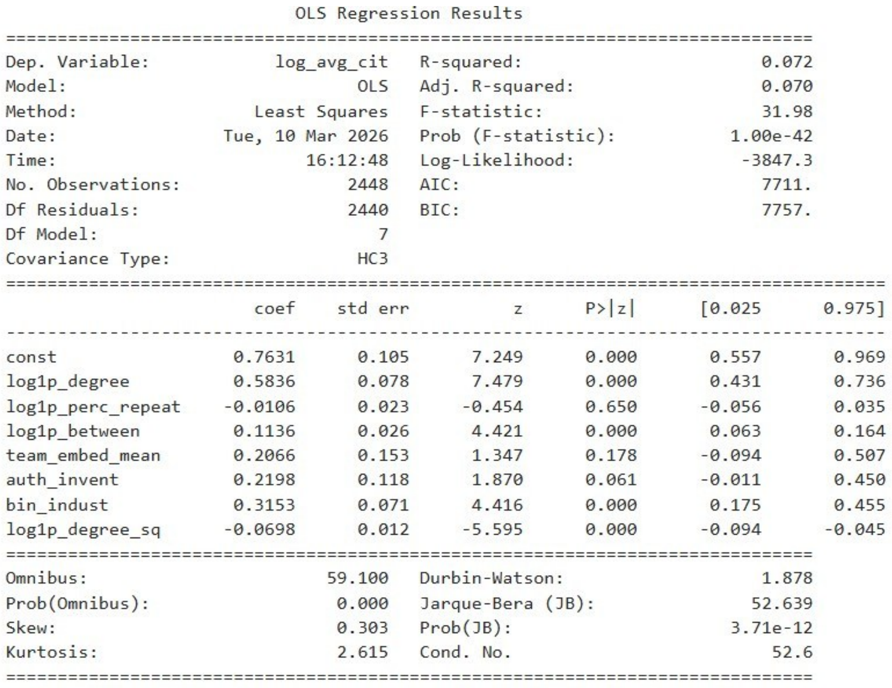
In contrast, the interaction between binary industry collaboration and degree centrality is statistically significant and positive in the regression results reported in Table 27. This finding indicates that the positive relationship between collaboration with industry and citation impact becomes stronger for researchers engaging in publications with more different colleagues. By comparison, the interaction between binary industry collaboration and betweenness centrality is statistically insignificant, suggesting that effect of industry collaboration on citation impact can not be significantly alter by betweenness centrality.

Overall, these results indicate that among the interaction terms examined, only the interaction between industry collaboration and degree centrality shows a statistically significant effect.



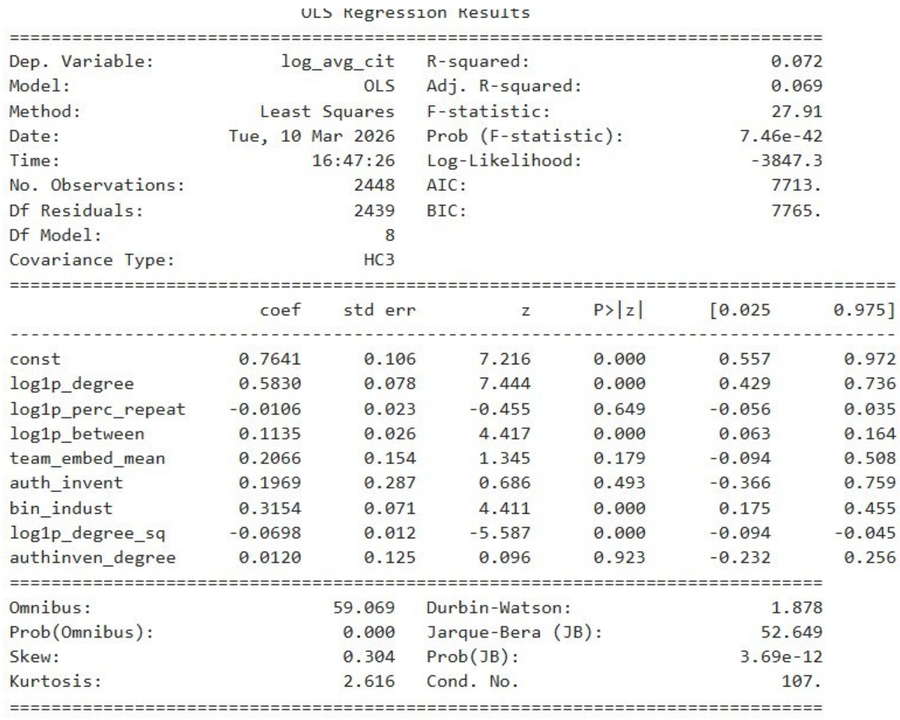
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.064
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     28.16
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           1.71e-37
Time:                        16:15:08   Log-Likelihood:                -3858.6
No. Observations:                2448   AIC:                             7733.
Df Residuals:                    2440   BIC:                             7780.
Df Model:                           7
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 1.1889      0.058     20.336      0.000       1.074       1.303
log1p_degree          0.2080      0.028      7.329      0.000       0.152       0.264
log1p_perc_repeat    -0.0027      0.024     -0.111      0.912      -0.050       0.045
log1p_between         0.1290      0.082      1.569      0.117      -0.032       0.290
team_embed_mean       0.0451      0.145      0.311      0.755      -0.239       0.329
auth_invent           0.2265      0.118      1.920      0.055      -0.005       0.458
bin_indust            0.2982      0.072      4.155      0.000       0.158       0.439
log1p_between_sq     -0.0052      0.015     -0.347      0.728      -0.035       0.024
==============================================================================
Omnibus:                       74.934   Durbin-Watson:                   1.880
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               60.913
Skew:                           0.309   Prob(JB):                     5.93e-14
Kurtosis:                       2.536   Cond. No.                         28.0
==============================================================================
```

Table 23. Quadratic specification for betweenness centrality



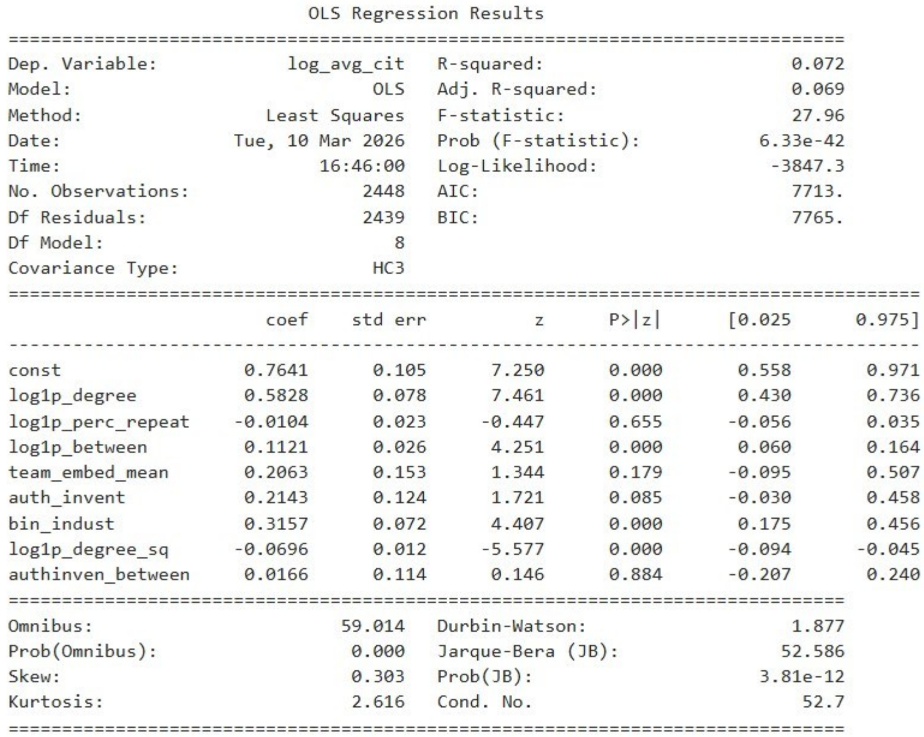
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.070
Method:                 Least Squares   F-statistic:                     31.98
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           1.00e-42
Time:                        16:12:48   Log-Likelihood:                -3847.3
No. Observations:                2448   AIC:                             7711.
Df Residuals:                    2440   BIC:                             7757.
Df Model:                           7
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.7631      0.105      7.249      0.000       0.557       0.969
log1p_degree          0.5836      0.078      7.479      0.000       0.431       0.736
log1p_perc_repeat    -0.0106      0.023     -0.454      0.650      -0.056       0.035
log1p_between         0.1136      0.026      4.421      0.000       0.063       0.164
team_embed_mean       0.2066      0.153      1.347      0.178      -0.094       0.507
auth_invent           0.2198      0.118      1.870      0.061      -0.011       0.450
bin_indust            0.3153      0.071      4.416      0.000       0.175       0.455
log1p_degree_sq      -0.0698      0.012     -5.595      0.000      -0.094      -0.045
==============================================================================
Omnibus:                       59.100   Durbin-Watson:                   1.878
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.639
Skew:                           0.303   Prob(JB):                     3.71e-12
Kurtosis:                       2.615   Cond. No.                         52.6
==============================================================================
```

Table 24. Quadratic specification for degree centrality



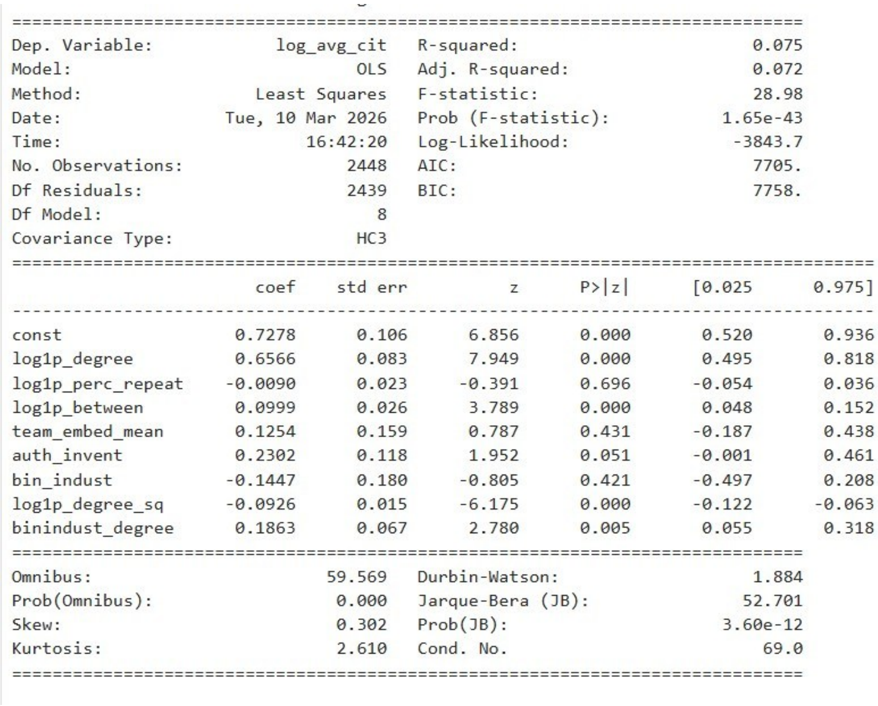
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.069
Method:                 Least Squares   F-statistic:                     27.91
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           7.46e-42
Time:                        16:47:26   Log-Likelihood:                -3847.3
No. Observations:                2448   AIC:                             7713.
Df Residuals:                    2439   BIC:                             7765.
Df Model:                           8
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.7641      0.106      7.216      0.000       0.557       0.972
log1p_degree          0.5830      0.078      7.444      0.000       0.429       0.736
log1p_perc_repeat    -0.0106      0.023     -0.455      0.649      -0.056       0.035
log1p_between         0.1135      0.026      4.417      0.000       0.063       0.164
team_embed_mean       0.2066      0.154      1.345      0.179      -0.094       0.508
auth_invent           0.1969      0.287      0.686      0.493      -0.366       0.759
bin_indust            0.3154      0.071      4.411      0.000       0.175       0.455
log1p_degree_sq      -0.0698      0.012     -5.587      0.000      -0.094      -0.045
authinven_degree      0.0120      0.125      0.096      0.923      -0.232       0.256
==============================================================================
Omnibus:                       59.069   Durbin-Watson:                   1.878
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.649
Skew:                           0.304   Prob(JB):                     3.69e-12
Kurtosis:                       2.616   Cond. No.                         107.
==============================================================================
```

Table 25. Interaction effect: author-inventor status and degree centrality



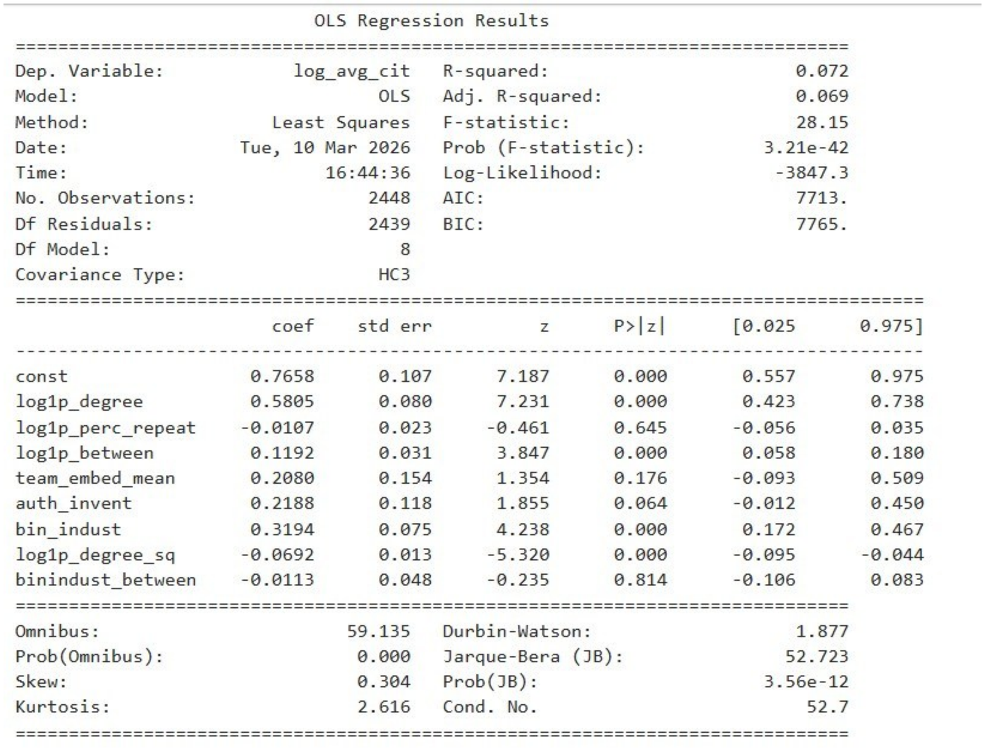
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.069
Method:                 Least Squares   F-statistic:                     27.96
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           6.33e-42
Time:                        16:46:00   Log-Likelihood:                -3847.3
No. Observations:                2448   AIC:                             7713.
Df Residuals:                    2439   BIC:                             7765.
Df Model:                           8
Covariance Type:                  HC3
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  0.7641      0.105      7.250      0.000       0.558       0.971
log1p_degree           0.5828      0.078      7.461      0.000       0.430       0.736
log1p_perc_repeat     -0.0104      0.023     -0.447      0.655      -0.056       0.035
log1p_between          0.1121      0.026      4.251      0.000       0.060       0.164
team_embed_mean        0.2063      0.153      1.344      0.179      -0.095       0.507
auth_invent            0.2143      0.124      1.721      0.085      -0.030       0.458
bin_indust             0.3157      0.072      4.407      0.000       0.175       0.456
log1p_degree_sq       -0.0696      0.012     -5.577      0.000      -0.094      -0.045
authinven_between      0.0166      0.114      0.146      0.884      -0.207       0.240
==============================================================================
Omnibus:                       59.014   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.586
Skew:                           0.303   Prob(JB):                     3.81e-12
Kurtosis:                       2.616   Cond. No.                         52.7
==============================================================================
```

Table 26. Interaction effect: author-inventor status and betweenness centrality



```
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.075
Model:                            OLS   Adj. R-squared:                  0.072
Method:                 Least Squares   F-statistic:                     28.98
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           1.65e-43
Time:                        16:42:20   Log-Likelihood:                -3843.7
No. Observations:                2448   AIC:                             7705.
Df Residuals:                    2439   BIC:                             7758.
Df Model:                           8
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.7278      0.106      6.856      0.000       0.520       0.936
log1p_degree          0.6566      0.083      7.949      0.000       0.495       0.818
log1p_perc_repeat    -0.0090      0.023     -0.391      0.696      -0.054       0.036
log1p_between         0.0999      0.026      3.789      0.000       0.048       0.152
team_embed_mean       0.1254      0.159      0.787      0.431      -0.187       0.438
auth_invent           0.2302      0.118      1.952      0.051      -0.001       0.461
bin_indust           -0.1447      0.180     -0.805      0.421      -0.497       0.208
log1p_degree_sq      -0.0926      0.015     -6.175      0.000      -0.122      -0.063
binindust_degree      0.1863      0.067      2.780      0.005       0.055       0.318
==============================================================================
Omnibus:                       59.569   Durbin-Watson:                   1.884
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.701
Skew:                           0.302   Prob(JB):                     3.60e-12
Kurtosis:                       2.610   Cond. No.                         69.0
==============================================================================
```

Table 27. Interaction effect: industry collaboration and degree centrality



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_avg_cit   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.069
Method:                 Least Squares   F-statistic:                     28.15
Date:                Tue, 10 Mar 2026   Prob (F-statistic):           3.21e-42
Time:                        16:44:36   Log-Likelihood:                -3847.3
No. Observations:                2448   AIC:                             7713.
Df Residuals:                    2439   BIC:                             7765.
Df Model:                           8
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.7658      0.107      7.187      0.000       0.557       0.975
log1p_degree          0.5805      0.080      7.231      0.000       0.423       0.738
log1p_perc_repeat    -0.0107      0.023     -0.461      0.645      -0.056       0.035
log1p_between         0.1192      0.031      3.847      0.000       0.058       0.180
team_embed_mean       0.2080      0.154      1.354      0.176      -0.093       0.509
auth_invent           0.2188      0.118      1.855      0.064      -0.012       0.450
bin_indust            0.3194      0.075      4.238      0.000       0.172       0.467
log1p_degree_sq      -0.0692      0.013     -5.320      0.000      -0.095      -0.044
binindust_between    -0.0113      0.048     -0.235      0.814      -0.106       0.083
==============================================================================
Omnibus:                       59.135   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.723
Skew:                           0.304   Prob(JB):                     3.56e-12
Kurtosis:                       2.616   Cond. No.                         52.7
==============================================================================
```

Table 28. Interaction effect: industry collaboration and betweenness centrality

4.4.8 Discussion of Citation Results

The final regression results provide several insights into the determinants of scientific productivity and citation impact. Overall, the findings highlight the important role of collaboration networks and industry engagement in shaping both the quantity and impact of scientific outputs.

First, the results indicate that collaboration structure plays an important role in explaining scientific productivity. In the productivity model, average group size shows a statistically significant negative relationship with the number of publications. This finding suggests that researchers who collaborate within larger teams may produce fewer publications individually. One possible explanation is that larger research teams may involve more complex coordination processes and division of labor, which can reduce the number of publications attributed to individual researchers. Previous studies have similarly suggested that larger collaboration teams may increase coordination costs and slow down publication output at the individual level (Wuchty et al., 2007).

By contrast, betweenness centrality shows a strong and statistically significant relationship with scientific productivity. The positive coefficient of betweenness centrality indicates that researchers occupying brokerage positions within collaboration networks tend to produce more publications. Scientists who connect otherwise disconnected groups may gain access to diverse knowledge sources and research opportunities, which can facilitate the generation of new research outputs. Similar positive associations between brokerage positions and research productivity have been reported in previous studies of scientific collaboration networks (Freeman, 1979; Newman, 2001; Abbasi et al., 2012).

At the same time, the quadratic term of betweenness centrality in the productivity model is negative and statistically significant, indicating an inverted-U shaped relationship between betweenness centrality and the number of publications. This finding suggests that occupying brokerage positions initially enhances research productivity, but the marginal benefits of increasing brokerage positions decline after a certain level. One possible explanation is that maintaining too many bridging relationships may increase coordination and communication costs, which can reduce the efficiency of research production. Similar nonlinear patterns between collaboration networks and research performance have been reported in previous studies (Abbasi et al., 2012; Uzzi et al., 2013).

The results further indicate that repeat collaboration intensity has a positive and statistically significant relationship with scientific productivity. Researchers who repeatedly collaborate with the same partners may benefit from improved coordination, trust, and familiarity, which can increase the efficiency of research production. Previous research has also found that repeated collaboration can facilitate knowledge sharing and improve research productivity (Uzzi et al., 2013).

In the citation model, degree centrality shows a positive and statistically significant relationship with citation impact. This finding suggests that researchers with a larger number of collaboration ties tend to receive higher citation counts. Larger collaboration networks can increase the visibility of research outputs and facilitate wider dissemination of scientific ideas. Similar positive relationships between collaboration intensity and citation impact have been widely documented in the scientometric literature (Wuchty et al., 2007; Abbasi et al., 2012).

However, the quadratic term of degree centrality is negative and statistically significant, indicating an inverted-U shaped relationship between collaboration intensity and citation impact. This pattern suggests that while increasing collaboration initially improves citation performance, the marginal benefits of additional collaborations decline after a certain level of network expansion. This result is consistent with the idea that very large collaboration networks may generate coordination costs and reduce the effectiveness of scientific collaboration (Abbasi et al., 2012; Uzzi et al., 2013).

Regarding industry engagement, the results indicate that industry collaboration is positively associated with both scientific productivity and citation impact. In particular, the percentage of industry collaborators in the productivity model and the binary industry collaboration indicator in the citation model both show positive and statistically significant effects. Industry collaboration may provide access to additional research resources, practical research problems, and broader knowledge networks, which can enhance both the productivity and impact of scientific research. Previous studies have similarly reported positive relationships between university–industry collaboration and scientific performance (Perkmann et al., 2013; Abramo et al., 2011).

The results also suggest that author–inventor status is positively associated with scientific productivity and citation impact, although the statistical significance of this relationship is moderate. Researchers engaged in both scientific publishing and patenting activities may benefit from knowledge spillovers between academic and technological domains. Such cross-domain knowledge flows can enhance both the novelty and applicability of research outputs (Meyer, 2006).

Finally, the interaction term between industry collaboration and degree centrality in the citation model is positive and statistically significant. This result indicates that the positive effect of industry collaboration on citation impact becomes stronger for researchers with higher levels of degree centrality. In other words, industry collaboration appears to generate greater citation benefits for scientists who maintain broader collaboration networks, likely because larger networks facilitate the diffusion and visibility of industry-related research outputs.

Chapter 5: Conclusion, Limitations, and Future Directions

5.1 Conclusion

The empirical results highlight several mechanisms through which collaboration networks and industry engagement influence both scientific productivity and citation impact. Overall, the findings suggest that researchers' positions within collaboration networks, together with their engagement with industry, play an important role in shaping scientific performance.

First, betweenness centrality shows a positive relationship with scientific productivity. Researchers occupying bridging positions within collaboration networks tend to produce more publications, likely because these positions allow access to diverse knowledge sources and collaboration opportunities. However, the negative and statistically significant quadratic term indicates an inverted-U shaped relationship between betweenness centrality and the number of publications. This suggests that while brokerage positions initially increase productivity, their marginal benefits decline beyond a certain level, possibly due to coordination and communication costs.

A similar nonlinear pattern appears in the citation model for degree centrality. Degree centrality positively influences citation impact, indicating that researchers with a larger number of collaboration ties tend to receive more citations. At the same time, the negative quadratic term suggests an inverted-U shaped relationship, meaning that the benefits of expanding collaboration networks eventually diminish as coordination becomes more complex.

The results also highlight the role of industry engagement in shaping research outcomes. In the productivity model, the percentage of industry collaborators has a positive and significant effect on the number of publications, suggesting that collaboration with industry can support research production through additional resources and practical research opportunities. Similarly, in the citation model, the binary indicator of industry collaboration shows a positive association with citation impact, indicating that researchers collaborating with industry tend to receive higher citation counts.

In addition, author–inventor status is positively associated with both scientific productivity and citation impact, suggesting that researchers engaged in both publishing and patenting activities may benefit from knowledge flows between scientific and technological domains.

Finally, the interaction analysis shows that the effect of industry collaboration on citation impact depends on researchers' collaboration networks. The positive and statistically significant interaction between binary of industry and degree centrality indicates that the citation benefits of industry collaboration become stronger for researchers with broader collaboration networks. This suggests that larger collaboration networks may enhance the visibility and diffusion of industry-related research outputs.

Overall, the findings indicate that scientific productivity and citation impact are shaped not only by collaboration networks themselves but also by how these networks interact with industry engagement and technological activities.

The findings of this study offer useful implications for policymakers, industry partners, and academic researchers. By highlighting the roles of collaboration networks and industry engagement in shaping both scientific productivity and citation impact, the results can help policymakers design policies that support effective university–industry collaboration and allocate research resources more strategically. Industry organizations may also benefit from these insights by strengthening partnerships with well-connected academic researchers. In addition, academic scientists can use these findings to develop more effective collaboration strategies and position themselves within research networks to improve both the quantity and impact of their research output.

5.2 Limitations

Despite its contributions, this study is subject to several limitations that should be acknowledged. First, the data were collected from Scopus for academic publications and Lens for patent records. While both are reputable sources, they may not offer full coverage of all Canadian researchers, particularly those working in interdisciplinary fields or whose affiliations and names are inconsistently recorded.

Another limitation concerns the identification of author–inventor overlaps across the Scopus and Lens datasets. Although the issue of initials was addressed by retrieving full author names from the original sources, the matching procedure relied primarily on exact name similarity. While differences in name order were allowed, cases where one record included a middle name and the other did not were excluded from the matching process. As a result, some true author–inventor pairs may not have been identified. In addition, individuals whose names have changed over time—for example, due to marriage or variations in name formatting—may not have been identified as matches and were thus excluded from the final list. Consequently, although the strict matching criteria reduce the likelihood of incorrect matches, the list of matched author-inventors may be incomplete and subject to potential classification errors.

Third, several unobserved factors that may influence research performance—such as grant funding availability, reputation, international collaboration intensity, or previous publication experience—were not captured in the available data. The omission of these latent variables may introduce potential bias in the interpretation of regression results.

Another limitation of this study lies in the measurement of science-industry interaction. We relied on two primary indicators—co-authorship with industrial-affiliated scientists and the presence of author-inventor overlap—to capture collaborative ties between academia and industry. While both approaches are widely used and provide valuable insights, they represent only a subset of the broader range of available methods. Other common indicators were not incorporated into this analysis. Moreover, none of these methods—ours included—can fully capture the complexity of science-industry collaboration on their own. Therefore, using multiple, complementary models in future research could enhance the reliability and depth of our understanding of how these interactions shape scientific performance.

Finally, the study relies on cross-sectional data from a fixed period (2015–2020) for publications). As such, it does not account for dynamic changes in collaboration patterns or citation behavior over time. Longitudinal data could provide deeper insights into the temporal aspects of scientific collaboration and performance.

5.3 Future Research Directions

Given the limitations of this study, several directions for future research are recommended.

As mentioned before, utilizing databases that provide more complete information on unique researcher identifiers (such as ORCID), institutional affiliations, and longitudinal records would enhance the accuracy of author-inventor disambiguation and improve the reliability of the analysis.

Also, future studies may consider additional approaches to measuring the interaction between science and industry beyond the ones applied in this research. For example, analyzing funding sources associated with scientific publications—such as industrial grants or public-private partnership funding—can offer insights into institutional ties between researchers and industry. Another important avenue is the use of mobility data, which tracks the movement of scientists between academic and industrial sectors over time. By identifying changes in institutional affiliation—such as a scientist shifting from a university position to a role in a private company or vice versa—researchers can better capture long-term collaborations, career-integrated partnerships, and the diffusion of knowledge across sectors. Mobility data can often be reconstructed from longitudinal affiliation records in publication metadata or professional CVs when available.

In addition, this study used a fixed time window to collect and analyze publication and patent data. While this approach is effective for capturing cross-sectional relationships, it may not fully reflect the temporal dynamics of collaboration. Future research could benefit from adopting a longer and more flexible time frame to observe evolving patterns of behavior. For example, applying a rolling or cumulative window over a decade or more may uncover gradual changes in collaborative structures, the sustainability of industry partnerships, and delayed effects of industrial involvement on academic performance. A non-fixed time frame would also allow researchers to capture the entry and exit of actors in collaborative networks and to analyze the continuity or discontinuity of key relationships.

Finally, incorporating additional explanatory variables may influence scientific productivity and citation impact. For example, researcher academic age, reflecting the length of a scientist's research career, may affect both publication output and accumulated citations. International collaboration intensity, representing the extent of cross-country research partnerships, may increase visibility and knowledge exchange. Institutional research prestige, referring to the scientific reputation and resources of the researcher's affiliated institution, may also influence research opportunities and impact. Grant funding availability may support research activities and enable larger or more ambitious projects. Finally, research field citation norms, which vary across disciplines, may affect how frequently publications are cited and therefore influence citation-based performance indicators.

RFERENCES

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars. Journal of Informetrics, 5(4), 594–607.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). Research productivity: Are higher academic ranks more productive than lower ones? Scientometrics, 88(3), 915–928.

Abramo, G., D'Angelo, C. A., & Zhang, L. (2019). The effects of collaboration on research performance. Scientometrics.

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. Science, 357(6351), 583–587.

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation. Administrative Science Quarterly, 45(3), 425–455.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality. Springer Handbook of Science and Technology Indicators.

Azagra-Caro, J. M., Fernandez-de-Lucio, I., & Gutierrez-Gracia, A. (2006). University patents: Output and input indicators. Research Evaluation, 15(2), 115–130.

Azoulay, P., Zivin, J. S. G., & Sampat, B. N. (2019). The diffusion of scientific knowledge. Nature Biotechnology, 37(5), 573–582.

Balconi, M., Brusoni, S., & Orsenigo, L. (2004). In defence of the linear model. Research Policy, 33(1), 77–89.

Bekkers, R., & Bodas Freitas, I. M. (2008). Analysing knowledge transfer channels. Research Policy, 37(10), 1837–1853.

Bianchini, S., et al. (2020). [Science-industry collaboration studies]. Research Policy.

Bikard, M., et al. (2019). Academic engagement and scientific output. Management Science.

Bonacich, P. (1987). Power and centrality. American Journal of Sociology, 92(5), 1170–1182.

Bornmann, L., & Daniel, H. D. (2008). What do citation count measure? Journal of Documentation, 64(1), 45–80.

Breschi, S., & Catalini, C. (2010). Tracing the links between science and technology. Research Policy.

Burt, R. S. (1992). Structural holes. Harvard University Press.

Burt, R. S. (2004). Structural holes and good ideas. American Journal of Sociology, 110(2), 349–399.

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data. Cambridge University Press.

Carayol, N., & Cassi, L. (2009). Who's who in science. Research Policy, 38(8), 1307–1321.

Chen, Y., Zhang, Y., & Fu, Y. (2017). Knowledge diffusion in academic networks. Scientometrics, 112(1), 441–463.

Chen, X. (2023). Citation networks and research evaluation. Scientometrics.

Civera, A., et al. (2020). Academic–industry collaboration. Research Policy.

Coleman, J. S. (1988). Social capital in the creation of human capital. American Journal of Sociology, 94, S95–S120.

Cook, K. S. (2014). Exchange and power in networks. Sociological Theory.

Crespi, G., & Geuna, A. (2008). Scientific production. Research Policy, 37(4), 565–579.

Criscuolo, P., & Verspagen, B. (2008). Patent citations. Research Policy, 37(10), 1890–1901.

Cummings, J. N., & Kiesler, S. (2005). Collaborative research. Social Studies of Science, 35(5), 703–722.

D'Angelo, C. A., et al. (2011). University–industry collaboration. Scientometrics.

Donner, P., et al. (2020). Industry publications. Research Evaluation.

Eslami, H., Beaudry, C., & Schiffauerova, A. (2013). Collaboration network performance. Scientometrics, 97(1), 99–119.

Etzkowitz, H., & Leydesdorff, L. (2000). Triple helix model. Research Policy, 29(2), 109–123.

Fini, R., et al. (2018). Academic entrepreneurship. Research Policy.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. Strategic Management Journal, 25(8–9), 909–928.

Fortunato, S., et al. (2018). Science of science. Science, 359(6379).

Freeman, L. C. (1979). Centrality in social networks. Social Networks, 1(3), 215–239.

Geuna, A., & Nesta, L. J. J. (2006). University patenting. Research Policy.

Glänzel, W., & Schubert, A. (2001). International co-authorship. Scientometrics, 50(2), 199–214.

Glänzel, W., & Schubert, A. (2004). Scientific networks. Springer.

Godin, B. (1996). Industrial publications. Research Policy, 25(4), 587–606.

Granovetter, M. (1973). The strength of weak ties. American Journal of Sociology, 78(6), 1360–1380.

Granovetter, M. (1985). Economic action and social structure. American Journal of Sociology, 91(3), 481–510.

Guerrero, M., et al. (2015). University technology transfer. Research Policy.

Gulati, R. (1995). Social structure and alliances. Administrative Science Quarterly, 40(4), 619–652.

Hicks, D., et al. (2015). Bibliometrics manifesto. Nature, 520, 429–431.

Hirsch, J. E. (2005). An index to quantify research output. PNAS, 102(46), 16569–16572.

Hottenrott, H., & Lawson, C. (2017). Research collaboration. Research Policy.

Hottenrott, H., & Thorwarth, S. (2011). Industry funding. Research Policy.

Ioanid, A., Scarlat, C., & Teodorescu, N. (2018). Network organization. Management & Marketing.

Jaffe, A. B. (1989). Academic research effects. American Economic Review.

Jefferson, O. A., et al. (2018). Patent citations. Nature Biotechnology.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning. Science.

Katz, J. S., & Martin, B. R. (1997). Research collaboration. Research Policy.

Ke, Q., et al. (2023). Citation network analysis. Nature Communications.

Larivière, V., et al. (2014). Team size and impact. JASIST.

Lee, S., & Bozeman, B. (2005). Collaboration and productivity. Social Studies of Science.

Leahey, E., et al. (2017). Interdisciplinarity. Administrative Science Quarterly.

Lissoni, F., et al. (2013). Scientific credit. Journal of Economic Behavior.

Magerman, T., et al. (2015). Patent mapping. Research Policy.

Marx, M., & Fuegi, A. (2020). Patent–publication links. Nature.

Merton, R. K. (1968). Matthew effect. Science.

Meyer, M. (2000). Patent citations. Research Policy.

Meyer, M. (2003). Academic inventors. Research Policy.

Moed, H. F. (2005). Citation analysis. Springer.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). Science–technology linkage. Research Policy.

Newman, M. E. J. (2001). Co-authorship networks. PNAS.

Opsahl, T., et al. (2010). Node centrality. Social Networks.

Perkmann, M., & Walsh, K. (2007). University–industry relations. IJMR.

Perkmann, M., et al. (2013). Academic engagement. Research Policy.

Phelps, C., et al. (2012). Knowledge networks. Academy of Management Journal.

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational collaboration. Administrative Science Quarterly.

Rafols, I., & Meyer, M. (2010). Interdisciplinarity. Research Policy.

Ranga, M., & Etzkowitz, H. (2013). Triple helix systems. Industry and Higher Education.

Reagans, R., & Zuckerman, E. (2001). Networks and productivity. Organization Science.

Rhoten, D., & Parker, A. (2004). Interdisciplinary research. Science.

Santoro, M. D., et al. (2019). Industry collaboration. Research Policy.

Schmoch, U. (1997). Industry collaboration. Scientometrics.

Secundo, G., et al. (2017). Technology transfer. Technological Forecasting.

Stephan, P. (1996). Economics of science. Journal of Economic Literature.

Tijssen, R. (2012). Industry collaboration. Research Evaluation.

Torvik, V. I., & Smalheiser, N. R. (2009). Name disambiguation. Bioinformatics.

Uzzi, B. (1997). Social structure and competition. Administrative Science Quarterly.

Uzzi, B., & Spiro, J. (2005). Creativity networks. American Journal of Sociology.

Van Raan, A. F. J. (2005). Citation analysis. Scientometrics.

Wagner, C. S., et al. (2011). International collaboration. Science.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure. Research Policy.

Waltman, L. (2016). Bibliometrics indicators. Journal of Informetrics.

Waltman, L., et al. (2011). Citation normalization. Journal of Informetrics.

Wang, J., et al. (2021). Collaboration networks. Scientometrics.

Watts, D. J., & Strogatz, S. H. (1998). Small-world networks. Nature.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). Team science. Science.

Zhang, L., et al. (2021). Collaboration and innovation. Research Policy.

Zhang, X., & Wu, Y. (2024). Citation impact analysis. Scientometrics.