SYM-AM-26-135

Excerpt from the Proceedings

of the

Twenty-Third Annual Acquisition Research Symposium and Innovation Summit

Volume III

"Accelerating Warfighting Capabilities"

Precision vs. Scope: A Methodological Framework for Linking Procurement and Patent Data in Defense Innovation Research

Published: April 30, 2026

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

The research presented in this report was supported by the Acquisition Research Program, Graduate School of Defense Management at the Naval Postgraduate School.

To request defense acquisition research, please contact:

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
E: arp@nps.edu
www.acquisitionresearch.net

Precision vs. Scope: A Methodological Framework for Linking Procurement and Patent Data in Defense Innovation Research

Capt Zachary O. Ryan, USAF—is a doctoral student in the Department of Systems Engineering and Management at Air Force Institute of Technology (AFIT). He received his BS in Technical Management from Embry–Riddle Aeronautical University and an MS in Program Management and Acquisitions from the AFIT. His research interests include program and acquisitions management, organizational behavior, and systems engineering. [Zachary.ryan.2@us.af.mil]

Dr. Jonathan D. Ritschel—is an Associate Professor of Cost Analysis in the Department of Systems Engineering and Management at AFIT. He received his BBA in Accountancy from the University of Notre Dame, an MS in Cost Analysis from AFIT, and a PhD in Economics from George Mason University. His research interests include public choice, the effects of acquisition reforms on cost growth in DoD weapon systems, and economic institutional analysis. [Jonathan.ritschel@afit.edu]

Dr. Edward D. White—is a Professor of Statistics in the Department of Mathematics and Statistics at AFIT. He received his BS in Mathematics from the University of Tampa, MAS from The Ohio State University, and PhD in Statistics from Texas A&M University. His primary research interests include statistical modeling, simulation, and data analytics. [Edward.white@us.af.mil]

Dr. John J. Elshaw—is an Associate Professor of Management in the Department of Systems Engineering and Management at AFIT. He received his BS in Accounting from the University of Akron, MBA from Regis University, and PhD in Management from Purdue University. His research interests include leadership, human–technology interaction, learning curves and the impact on acquisition, workgroup and team processes within organizations, social networks, organizational climate and culture, cross-cultural leadership and communication, cognition and emotions, motivation, human performance, organizational trust, performance, and commitment. [john.elshaw.1@us.af.mil]

Abstract

Understanding innovation within the U.S. Defense Industrial Base (DIB) is a critical objective of researchers seeking to inform defense procurement policy, yet empirically linking government funding to innovation indicators remains challenging. This paper addresses the core methodological problem of connecting government procurement data with patent data, two systems that are not naturally aligned. We propose a matching framework centered on two distinct data-matching strategies: award linkage (L_A), which matches patents to specific government awards, and organization linkage (L_O), which matches patent assignees to contractor entities. We argue that these strategies present researchers with a structural trade-off. Award linkage offers high precision and defense specificity but suffers from narrow scope and under-classification of innovative activity, while organization linkage provides broad scope to analyze firm-level capacity and knowledge spillovers but risks over-classification by including non–defense-related patents. We develop and validate our matching framework by reviewing how recent literature has navigated data matching challenges, formalizing the most prevalent linkage strategies, and then presenting a practical demonstration of the proposed L_A – L_O framework by analyzing the patenting activity of the government's SBIR programs over the past decade.

Introduction

In an era of renewed strategic competition, U.S. defense industrial policy has pivoted toward accelerating technological advantage through a deliberate revitalization of industrial capacity. The Department of Defense's (DoD's) National Defense Industrial Strategy (NDIS; Hicks, 2023) calls for modernization of the defense industrial ecosystem to deter and defeat peer adversaries. To that end, the DoD has stated its intent to leverage the breadth of the U.S. economy with a specific focus on small businesses, innovation centers, and nontraditional suppliers to drive innovation and generate cost-effective solutions to defense-specific problems. Recent DoD policy discourse further sharpens the NDIS's technological imperative by identifying innovation as an operational outcome and emphasizing that the Defense Innovation Ecosystem should be deliberately structured to drive technological advantage.

The memorandum Transforming the Defense Innovation Ecosystem to Accelerate Warfighting Advantage (Secretary of War, 2026), outlines a high-level structural framework that aligns innovation efforts towards three innovative outcomes: technology, product, and operational capability. This outcome-driven framing is consistent with broadly accepted definitions of innovation, which, according to the Oslo Manual (2018), requires both a new or improved product or process and the implementation of that product or process in a way that makes it available and useful to end users or units. Succinctly, innovation is defined by both novel invention and active use.

This outcome-driven definition of the innovation process drives a familiar measurement tension. Innovation processes are nonlinear, and their outputs are difficult to observe directly. As a result, empirical research often proceeds by decomposing the process into individual components using established approaches grounded in the economic, management, and social sciences. Defense innovation research extends these tools and indicators to the Defense Industrial Base (DIB), a domain shaped by constraints stemming from the distinctive overlap of the corporate and government sectors (OECD/Eurostat, 2018).

Not all innovation indicators are equally informative under these constraints. Government contract records and R&D accounts provide a trove of input data through recorded appropriations and obligations (CRS, 2023); however, these same data do not explicitly capture realized effort or outcomes. Conversely, traditional business performance measures, such as firm growth or commercial success, are not necessarily aligned with the government program objectives, and they can be difficult to interpret when considering the complex contracting arrangements and administrative discontinuities that tie government and corporate systems (Hart et al., 1997; Rascona et al., 2023). In light of these considerations, defense innovation researchers have gravitated toward patents as a primary external indicator, as they are clearly defined and widely observed, provide explicit ties to inventive activity, are precisely time-stamped, and are widely used by economic researchers seeking to understand technical change (Griliches, 1990; Jaffe & Trajtenberg, 2002). This choice drives a central challenge that defense innovation researchers must resolve prior to analysis: establishing credible and defensible links between entities of interest and patents, i.e., resolving entities across administrative systems (Hall et al., 2001).

This paper contributes to the defense innovation discourse by reviewing existing defense innovation literature and extracting the implicit linkage assumptions leveraged in their designs, characterizing the two dominant linkage strategies—awards to patents (L_A) and organizations to patents (L_O)—and their failure modes, and developing a methodology to formally codify a linkage framework to guide researchers. It concludes with a practical demonstration of the framework by analyzing SBIR-associated patenting from 2014 to 2024 to illustrate how the L_A – L_O framework can be used to produce transparent and interpretable inferences. Our aim is to provide a reproducible foundation that makes linkage logic explicit so that researchers and policy-makers can interpret results appropriately.

Literature Review

This section reviews recent defense innovation literature to identify the most common linkage strategies employed by researchers. Using patent data to bridge the government–corporate divide and evaluate the DIB as an innovation system raises a foundational empirical challenge: patents and government accounting data are organized around different units, definitions, and identifiers. Government accounting data describe obligations, vendors, and programs through contracts and awards; patent data describe inventions, assignees, and knowledge linkages. Bridging these systems, therefore, requires careful choices about how to construct matches. Two primary data unification approaches have come to dominate the recent literature. The first is award-based linkage, which we define as L_A, which ties patents to specific contracts via funding acknowledgments and government-interest statements. The second is organization-based linkage, which we define as L_O, which matches assignees to vendors and corporate entities. Each approach provides insights into different policy questions and carries with it competing measurement risks.

Studies relying primarily on organization linkage L_O typically narrow organizational scope to reduce misclassification. Howell (2017) restricted attention to SBIR participants, where firms are by definition small and manually verifiable, allowing for accurate curation of the linkage set. Raiteri et al. (2018) leveraged L_O for broad organizational attribution but then applies an award-based filter (L_A) to classify defense-related patents, thereby avoiding overcounting of non-defense works. Howell et al. (2017, 2021) explored SBIR innovation through L_O, and in their later work, they expanded their work to include large primes. Fischer (2023) also employed L_O but limited organizational scope to entities engaged in GAO bid protests, implicitly filtering his set to defense-relevant firms. Each of these studies acknowledged the risk of over-attribution inherent in L_O and addressed it through controls that constrain the organizational set included within the L_O boundary.

Conversely, studies anchored in award linkage L_A emphasize precision but accept that coverage may be incomplete. Bruce et al. (2018) used L_A as their primary linkage, with L_O included as a robustness check to capture broader organizational effects. De Rassenfosse et al. (2019) restricted themselves to explicit award–patent identifiers, prioritizing clean attribution at the cost of scope. Decarolis et al. (2021) adopted the same logic, focusing on award-linked patents while leaving contractor attribution implicit. Giuffrida and Raiteri (2023) followed a tightly scoped L_A design, again prioritizing certainty of attribution over breadth. Bottai et al. (2025) similarly adopted L_A as their baseline, though they supplemented with organizational linkages from U.S. Patent and Trademark Office (USPTO) data to enable commercialization analysis.

The primary objective of these papers is not to establish a generalized data linkage method, and as a result, linkage strategies are often implied rather than declared. The mitigation tactics employed to handle data challenges are embedded in design choices rather than articulated directly. Nevertheless, researchers consistently implement safeguards to control for the precision-scope trade-off implicit in the two dominant strategies. Table 1 summarizes these patterns by classifying recent defense-specific contributions according to their primary linkage strategy and the mitigation strategies each employed to address misattribution, coverage gaps, or over- or under-classification.

Table 1. Classification of Recent Defense-Innovation Literature Matching USPTO Data to Government Contract Data

Study	**Linkage**	**Mitigation Strategy**
Financing Innovation (Howell, 2017)	L_O	Narrow organization scope (SBIR orgs)
A Time to Nourish (Raiteri et al., 2018)	L_O	LO for org attribution, LA for defense classification
Public Contracting (Bruce et al., 2018)	L_A	LA for primary measure, LO as control variable

Procurement of Innovation (de Rassenfosse et al., 2019)	L_A	Explicit award – patent linkages
Opening Up Military Innovation (Howell et al., 2021)	L_O	Narrow organizations scope (SBIR orgs & large primes)
Buyers' Role (Decarolis et al., 2021)	L_A	Scoped to award-patent link, contractor attribution undefined
The Missing Link (Fischer, 2023)	L_O	Narrow organization scope (GAO protests)
Bureaucratic Frictions (Giuffrida & Raiteri, 2023)	L_A	Tightly scoped to award-patent link
Measuring Commercialization (Bottai et al., 2025)	L_A	Tightly scoped to award-patent link, org linkage from USPTO

Note. Researchers have consistently confronted the structural trade-off between L_A and L_O. What varies across studies is how they mitigate the limitations of their chosen strategy.

The Data Landscape: Two Data Systems That Do Not Naturally Align

Having identified the most common linkage approaches from recent literature, we now turn our attention toward two of the most commonly used award and patent datasets to understand their structural composition and limitations. The first system discussed is the award layer, commonly retrieved from FPDS or USASpending, which captures procurement and assistance actions of the U.S. federal government. In this system, the atomic unit is the award, which is uniquely identified by a Procurement Instrument Identifier (PIID), a Federal Assistance Identifier (FAIN), or a Unique Record Identifier (URI). Awards capture detailed metadata about the award itself, the recipient of the award, and its funding and managing entities. Each award record is associated with a single prime recipient organization whose identity is specified via a Unique Entity Identifier (UEI). UEIs are drawn from SAM.gov registration data and are intended to represent distinct legal entities. In practice, however, defense-sector organizational structures are often shaped by business strategies, tax considerations, and partnering arrangements that can result in complex corporate hierarchies and multiple UEIs within the same corporate family (Defense Contract Audit Agency, 2019).

To illustrate the complexity, Table 2 reports the number of active UEIs (defined as UEIs that either directly or through a child received an award action) associated with the five largest U.S. defense contractors between 2014 to 2024. For each contractor, the table also reports the number of parent UEIs, defined here as UEI's that did not have a defined parent within the USASpending dataset. Table 2 provides a glimpse into the complex legal structures maintained by the largest defense organizations and highlights why aggregating awards under organizations is nontrivial.

Table 2. Active and Parent UEIs for the Five Largest Cefense Contractors, 2014–2024

Prime Contractor [search string]	**Total UEIs**	**Total Parent UEIs**
Lockheed Martin *[Lockheed]*	140	36
Boeing *[Boeing]*	76	12
Raytheon *[Raytheon]*	119	33
General Dynamics *[General Dynamics]*	120	30
Northrop Grumman *[Northrop]*	139	46

Note. Data retrieved from USASpending dataset.

Although awards provide direct attribution to a single UEI, the award layer also contains internal structures that can complicate the interpretation of "award" as a single analytical unit. For example, some award vehicles such as Indefinite Delivery Vehicles (IDVs) and Indefinite Delivery Contracts (IDCs) may have uniquely identified child awards that are issued to contractors that differ from the primary vehicle holder. Additionally, awards are modified by transactions that record the sequence of contractual changes over time and can diverge from award summaries. Taken together, awards and organizations comprise the primary entities of the USASpending award layer.

The second data system is the patent data layer, which captures intellectual property filings with the USPTO. In this system, the atomic unit is the patent, a legal object that confers a set of intellectual property rights on its owner. Patent records also contain detailed metadata, which includes inventor information, interested parties, filing and grant dates, technology classifications, and assignees. The assignee is the entity, whether an individual, corporation, university, or government agency, that holds legal ownership of the invention at the time of filing or subsequent assignment.

Patents are legal property rights, not procurement or project records. Simplified, patents are property, whereas awards are contracts. As such, the patent assignee field reflects the legal owner of the invention, which in many cases is a corporate parent or holding company rather than the operational unit that executed the inventive activity. In some organizations, intellectual property is managed centrally through dedicated legal offices, further decoupling patents from the specific subsidiaries or business units that generated the underlying innovation.

Taken together, patents and their assignees form the primary entities of the patent data layer. Patents can be mapped to these assignees with reasonable coherence through modern disambiguation processes; however, the absence of standardized identifier fields at the point of data entry, combined with the diversity of corporate intellectual property management strategies, limits the ability to infer clear and comprehensive organizational hierarchies from USPTO data alone. Constructing such hierarchies requires integrating additional external datasets, such as commercial corporate family databases or manually curated organizational mappings.

In summary, the USASpending award layer and the USPTO patent layer each provide rich, internally consistent views into distinct organizational systems that, if connected, could yield valuable insights into the innovation process. Linking them, however, is challenging. Each layer was developed for a specific purpose, employing different data management strategies, and as a result they do not naturally align. Awards are contractual units tied to a single recorded UEI, while patents are legal property rights tied to an assignee that may be disambiguated into a stable organizational entity. No standardized cross-layer key exists. Reconciling these systems therefore requires explicit linkage strategies, each of which rests on assumptions about the relationships between awards, patents, and their associated organizations. The next section formalizes the two primary linkage strategies into the $L_A - L_O$ framework.

The $L_A - L_O$ Framework

We have established that the USASpending award data and the USPTO patent layer define two internally complex and distinct relational systems. We now formally define them. Let $G_A = (V_A, E_A)$ denote the award layer where V_A consists of awards and organizations and E_A consists of contractual and hierarchical relationships (award–organization attribution, parent–child organizational ties, and award–award derivations). Similarly, let $G_P = (V_P, E_P)$ denote the patent layer, where V_P consists of patents and disambiguated assignees, and E_P consists of intellectual property relationships (patent–assignee attribution, patent–patent citations). By construction, G_A and G_P have no naturally occurring cross-layer edges; all such edges must be inferred or derived.

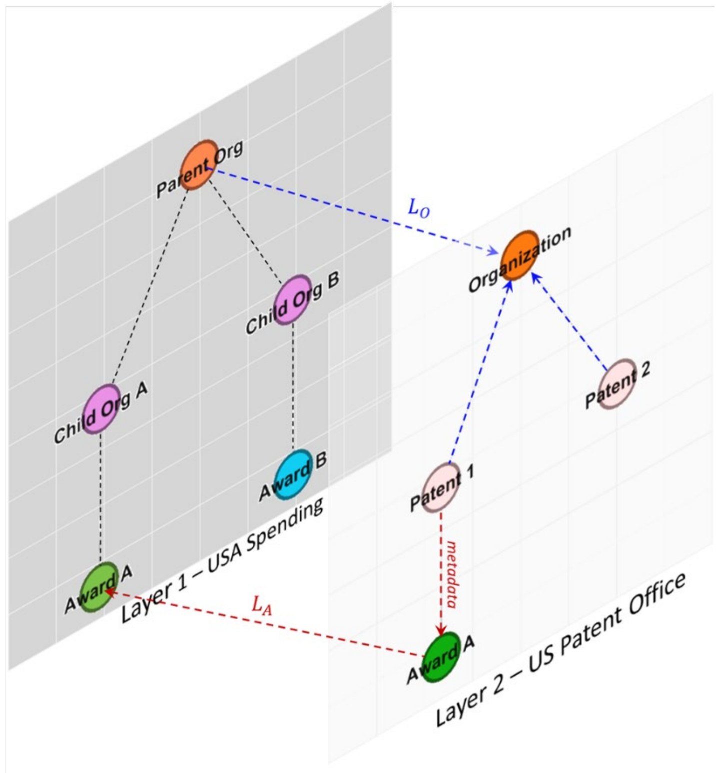

Figure 1. Comparative representation of award linkage (red) and organization linkage (blue) between the USASpending award layer (left) and the USPTO patent layer (right). Award linkage maximizes precision while organizational linkages maximize scope.

Let L be a linkage mapping that defines a subset of $V_A \times V_P$ representing hypothesized equivalences or associations between award-layer entities and patent-layer entities. Two canonical linkage strategies, award linkage and organization linkage, form the basis for most existing work in the literature. To illustrate, Figure 1 provides a visual depiction of both strategies—award linkage (L_A) in red and organizational linkage in blue (L_O)—illustrating how cross-layer edges can be established. These two approaches, while appearing similar at first glance, have fundamental trade-offs that must be formally defined.

Strategy 1: Award Linkage

Let A denote the set of awards in the USASpending layer and P the set of patents in the USPTO layer. Award linkage seeks to establish a mapping $L_A \subseteq A \times P$ such that $(a, p) \in L_A$ if and only if patent p can be directly attributed to award a via explicit identifiers such as by a PIID derived from the government interest statement or related metadata. This mapping provides a high-precision measure of which awards result in identifiable innovative outputs, since each link is grounded in an explicit contractual reference.

In practice, however, coverage is sharply constrained. Even when an R&D award results in a project culminating in a patent, the PIID may be omitted from the patent record, which results in its exclusion from the dataset. This limitation is both structural and a matter of data quality since reporting rules allow for some omissions. Of particular interest to researchers are those privately funded patents which may be closely related to award activity which are not subject to reporting rules (Rights to Inventions, 2023). As a result, the award linkage strategy inherently underestimates defense innovation outputs.

A second important but more nuanced limitation of L_A prevents the analyst from guaranteeing a clear organizational linkage from $G_P$to G_A, due to the differing characterizations of the organization entity in the two datasets. This nuance is best illustrated through a concrete example. Let $p_1 \in V_P$ denote a patent in the USPTO data that contains a valid government interest statement referencing an award identifier PIID. In the patent layer, p_1 is attributed to an assignee organization $o_{p1} \in O_P$. Because p_1 cites a PIID, a valid pair $(a_1, p_1) \in L_A$ can be established where $a_1 \in A \subset V_A$ denotes the corresponding award record in the USASpending layer. From this award record, one can further identify the recorded prime contractor $o_{c1} \in O_A$ (identifiable via UEI).

A naive inference would be to assume equivalence between the two organizational nodes $o_{p1} = o_{c1}$, implying that the patent's assignee is the same as the prime award recipient. However, this equivalence does not necessarily hold. Consider the case where p_1 was filed by a subcontractor $o_{c2} \in O_A$ performing work under a_{c1} on behalf of the prime o_{c1}. In this case, the cross-layer edge $(a_{c1}, p_1) \in L_A$ would erroneously induce the organizational inference $o_{p1} = o_{c1}$, even though the inventive activity originated from o_{c2}. This failure mode is easily illustrated by returning non-matching organizations via the L_A linkage. Consider $L_A(\mathrm{p}, \mathrm{a})\ where\ L_A(NNJ06TA25C, 10161690)$ resolves $(o\_a, o\_p\ )\ to\ (Lockheed\ Martin, Hamilton\ Sundstrand)$. Clearly $o_a \neq o_p$.

This failure mode generalizes. Because procurement records in G_A only record the prime contractor's UEI, the award–patent mapping L_A cannot distinguish between prime, subcontractor, or passthrough entities. In nonstandard contracting scenarios such as in cases of joint ventures or nested parent–subsidiary structures, the UEI used to execute a contract may not correspond to the true IP-generating entity in G_P. Thus, while L_A provides high-precision award-to-patent matches, it cannot guarantee correct organizational attribution, limiting its effectiveness for organizational or system level analyses.

Strategy 2: Organization Linkage

Let O_A be the set of organizational entities in the USASpending layer and O_P the set in the USPTO layer after disambiguation. Organization linkage defines a mapping $L_O \subseteq O_A \times O_P$ such that $(o_a, o_p) \in L_O$ if and only if the entities o_a and o_p are determined via a formally defined entity resolution process to represent the same underlying legal entity. Once established, this mapping allows the inheritance of all patents associated with o_p to o_a enabling the measurement of organizational and system-level innovation patterns. Formally, if an award-layer organization and a patent-layer assignee are matched, such that $\left(o_a, o_p\right) \in L_O$ then any patent p associated with the assignee, where $\left(o_p, p\right) \in E_P$ is inferred to be inherited by o_a.

A natural inference is to assume that all patents inherited through L_O are defense-related. This assumption is not unreasonable when one thinks of the largest prime contractors, such as Lockheed Martin, Raytheon, or Northrop Grumman, where the majority of patenting activity arises directly from defense-related R&D. In these cases, the mapping between $\left(o_a, o_p\right)$ and defense-relevant patents may appear valid.

However, this reasoning fails once nontraditional contractors are considered. Suppose $o_{p3} \in O_P$ corresponds to Apple, Inc., matched to $o_{a3} \in O_A$. By inheritance, all patents owned by Apple to include those related to iPhones, cloud services, and consumer electronics are transferred to the defense-linked set of o_{a3}. While some fraction of Apple's patenting may stem from defense-relevant projects, the overwhelming majority do not. The broad assumption that every inherited patent is defense-related therefore breaks down. This risks a structural over-classification failure as the assumption that the firm's entire patent portfolio, $\{p|(o_{p3}, p) \in E_P\}$, is a subset of defense-related patents, $(P_{defense})$ does not hold.

This misattribution is particularly problematic for diversified technology firms (e.g., Apple, Google, Amazon) or large firms with centralized IP management structures where large portfolios of unrelated patents are collapsed under a single corporate entity. The result is upward bias and a loss of subunit detail, the opposite of the under-attribution risk inherent in L_A.

Despite these risks, L_O provides analytical leverage unavailable to L_A. Because it links organizations rather than individual awards, it enables the study of innovation networks, spillover flows, cumulative knowledge growth, and organizational capacity. These are phenomena that cannot be measured through direct award-patent linkages alone. The trade-off between scope and precision, however, is unavoidable.

The Measurement Trade-Off

Under current public data availability and reporting practices, no single linkage function L^* can simultaneously achieve the contract-level precision of L_A (award linkage) and the system-level scope of L_O (organization linkage). This trade-off is structural and cannot be resolved through matching procedures. Consequently, the choice of linkage must be primarily driven by the innovation question under study. Questions that focus on organizational capacity, knowledge exchange, or spillover effects require L_O and subsequently must account for the inclusion of unrelated patents. On the other hand, questions focusing on direct inventive outputs as a result of government funding require L_A and, as such, must accept incomplete coverage and the loss of broader organizational context.

The contrast between the two linking strategies is most evident when comparing their operational characteristics (Table 3). Award linkage anchors measurements at the level of individual awards, producing high precision and strong defense specificity but only narrow scope. Organization linkage, in turn, shifts the unit of analysis to the firm, trading away defense specificity and precision in order to capture spillovers, network ties, and firm-level capacity. Because these differences are mechanical consequences of the chosen boundary definition, empirical results should be interpreted as conditional on linkage choice rather than as competing estimates of a single underlying quantity. The next section demonstrates this empirically by applying the $L_A - L_O$ framework to the SBIR ecosystem.

Table 3: Comparative Attributes of Award Linkage and Organizational Linkage

Dimension	**Award Linkage L_A**	**Organization Linkage L_O**
Primary Unit of Linkage	Award - Patent	Assignee - Organization
Linking Method	Explicit (PIID)	Formal Entity Resolution
Precision	High	Low
Scope	Narrow	Broad
Defense Specificity	High	Low
Primary Unit of Measure	Award	Organization
Key Limitation	Cannot capture general inventive activity	Introduces private inventive activity

L_A – L_O in Practice: SBIR Analysis (2014–2024)

To illustrate the practical trade-offs and implementation challenges inherent in both the organizational linkage L_O and award linkage L_A frameworks, we apply each mapping to USASpending and USPTO data to measure patenting activity associated with the DoD's SBIR programs. Having established the conceptual distinctions between award and organization level linkages, we now describe an empirical construction of these mappings designed for large-scale administrative government data, where reporting variations and record inconsistencies have introduced noise that must be handled explicitly. The following section describes the formally defined processes used to derive both L_O and L_A mappings for the SBIR data.

L_O – Organizational Linkage Process

This section describes the formally defined entity resolution process utilized to derive the L_O mapping. As established, let O_A be the set of organizational entities in the USASpending layer and O_P the set in the USPTO layer after disambiguation. The organizational linkage process progresses through a multistage approach consisting of four stages. Figure 2 depicts the primary stages of the process and its major functions adopted to a PRISMA style diagram. The process is designed to balance the complexities inherent to string-matching with the size of the dataset by leveraging the general framework applied by the PatentsView team (Monath et al., 2021) to disambiguate organizations within the USPTO patent database.

The process begins by filtering both datasets to identify the unique entities active within a specified timeframe. For the USASpending data, we identified 598K active UEI's between 2014 and 2025 by searching FPDS and eSRS transactions. We also identified 279K unique assignee IDs from the USPTO PatentsView disambiguated database active between 2010 and 2025. Restricting the USASpending data to 2014 to 2025 is necessary due to sub-contract reporting changes that occurred in 2013, while extending the patent organization window to include 2010 to 2025 allows for the capture of pre-award patenting activity. By isolating these active populations, we define the initial list of entities for the subsequent screening and eligibility phases.

The screening stage reduces the massive cross-product $O_A \times O_P$ into a computationally manageable set of high-probability pairs. To achieve this, we employ the 4-character canopy blocking approach leveraged by Monath et al. (2021). This blocking process filters the 167 billion candidate pairs down to approximately 1.5 billion pairs that warrant further algorithmic processing.

The eligibility stage is the most complex. It begins by processing the candidate *uei – assignee_id* pairs using a character n-gram model trained on the entire dataset using two- to four-character n-grams. Each pair is assigned a cosine similarity score representing how similar two organization names are compared to all other organization names (TF-IDF weights). Pairs with a score exceeding .6 are captured to further reduce the number of possible pairs. This yields 1.4M pairs, which are than processed with a tailored two-pass algorithm.

The first pass of the algorithm identifies matches by establishing links where names are identical, have minor spelling variations, or use recognizable acronyms. This process yields around 27K pairs which we label Pass 1. The second pass assesses entities that were not matched in the first pass. This pass identifies pairs whose cosine similarity score exceeds a threshold of .95, which was determined after performing a sensitivity analysis by running the L_O pipeline at different thresholds. Figure 3 summarizes the results of both passes illustrating how the final output of 28K pairs is heavily dependent on the chosen threshold defined in the second pass. The .95 threshold was selected because it minimized ties between pairs. When ties did occur, they were resolved using the geographic data associated with the entities and, if necessary, a temporal overlap of their patenting and award activity to select the most likely candidate pairs.

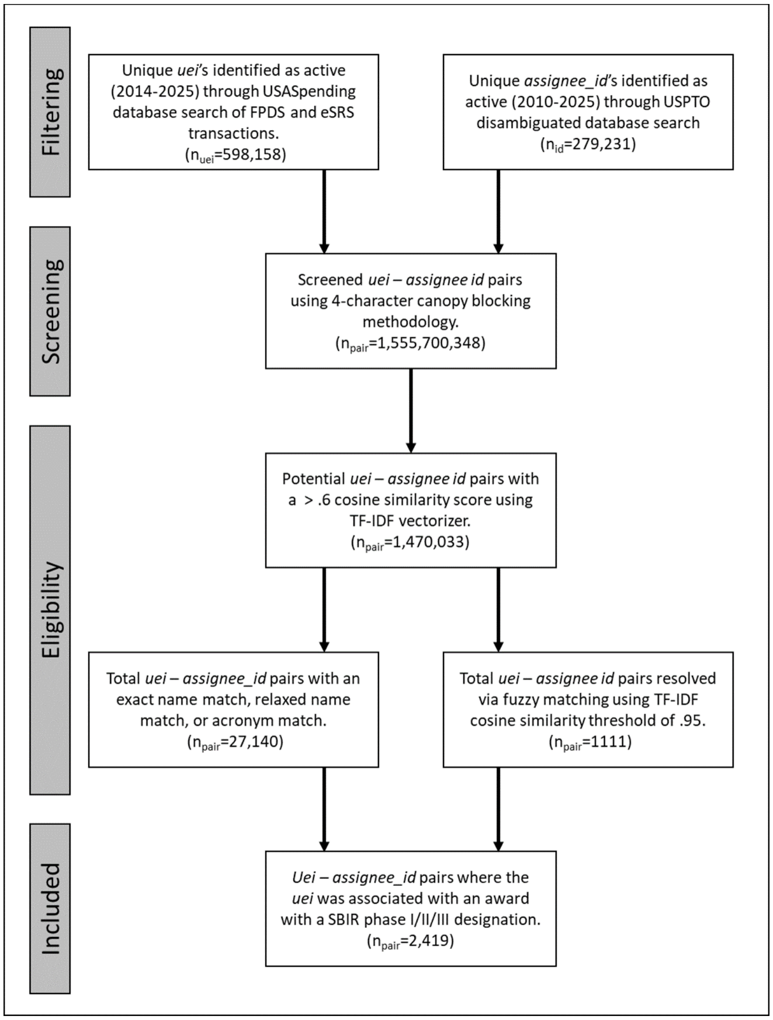

Figure 2: Organizational linkage entity resolution (ER) process. The Lo process utilized a four-stage pipeline to derive cross-layer matches between the USAspending layer and the USPTO patent layer. The ER process reduces the number of possible pairs from 1.5B to 3,222 using the computationally efficient approach applied by Monath et al. (2021).

The inclusion phase identifies organizations associated with the SBIR program from the pool of resolved pairs. Pairs were included in the final dataset if their *uei* was associated with an award designated as a SBIR phase I/II/III. The formal logic driving the entity resolution process is defined as follows.

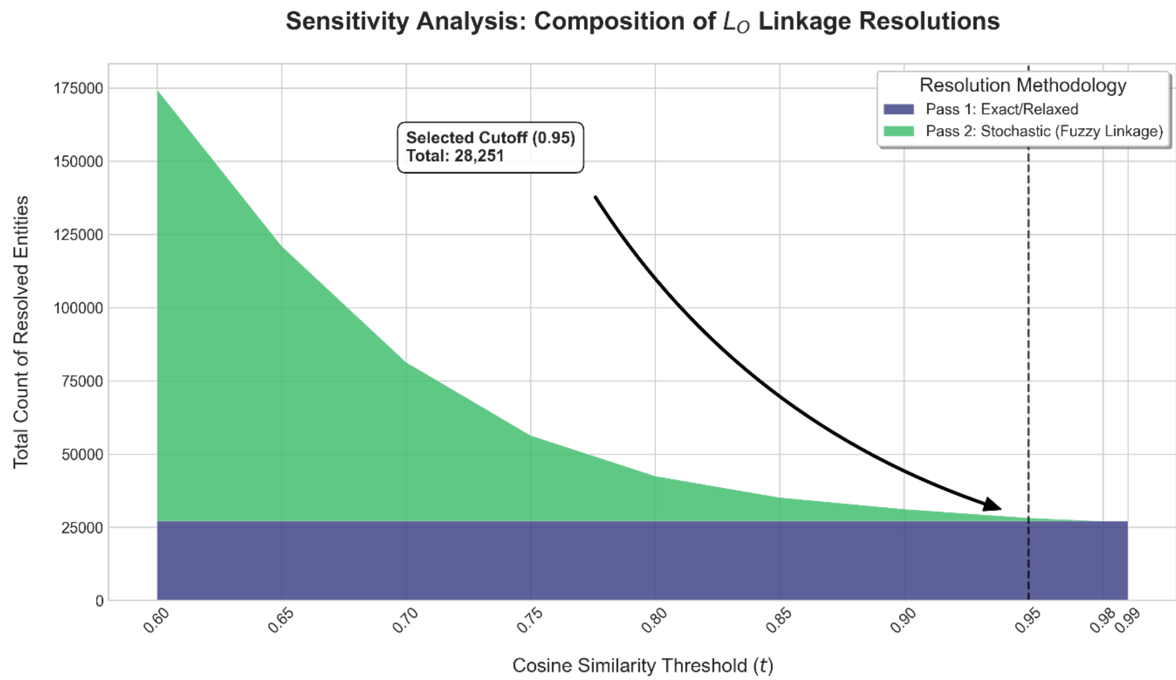

Figure 3: Sensitivity analysis of the Lo cosine-similarity cutoff. A fuzzy matching strategy was leveraged to resolve similar uei - assignee_id matches. Pass 1 matches are stable across thresholds while pass 2 matches approach zero as the threshold is tightened. A thigh precision threshold of .95 was selected because it limited the number of ties in the set.

The USASpending dataset does not consistently resolve to the parent corporate entity, and the USPTO dataset has been resolved to a singular *assignee_id* via disambiguation; therefore, we constraint the match from *uei* to *assignee_id* using a many to one relationship. Formally:

$$\forall o_a \in O_A: \left|\left\{o_p \in O_P: \left(o_a, o_p\right) \in \ L_O\right\}\right| \leq 1$$

$$\forall o_p \in O_P: \left|\left\{o_a \in O_A: \left(o_a, o_p\right) \in \ L_O\right\}\right| \ is\ unrestricted$$

Pass 1 – Exact and Relaxed Name Matching

For each $\left(o_a, o_p\right) \in O_A \times \ O_P$ let $n_a(o_a)$ and $n_p(o_p)$ denote their cleaned names and let $g\left(o_a, o_p\right) \in [0,2]$ denote a score representing geographic similarity. Define the name-match function:

$$F(n_a, n_p) = \{exact\left(n_a(o_a), n_p(o_p)\right) = 1\ or\ relaxed\left(n_a(o_a), n_p(o_p)\right) = 1\ or\ acronym\left(n_a(o_a), n_p(o_p)\right) = 1\}.$$

For each $o_a \in O_A$ define the candidate set as:

$$C_a(o_a) = \{o_p \in O_P: F\left(n_a, n_p\right) = 1\}.$$

If $|C_a(o_a)| = 1$ let $o_p^*(o_a)$ represent the unique element of $C_a(o_a)$ and include $\left(o_a, o_p^*(o_a)\right) \in \ L_O$. If $|C_a(o_a)| > 1$, let $o_p^*(o_a)$ represent the element from $C_a(o_a)$ with the greatest $g\left(o_a, o_p\right)$ and include $\left(o_a, o_p^*(o_a)\right) \in \ L_O$.

Pass 2 – Cosine Similarity Threshold Matching

Let $s(o_a, o_p) \in [0,1]$ denote the TF-IDF cosine similarity between $n_a(o_a)$ and $n_p(o_p)$ and set a fixed threshold $t \in (0,1)$. For each unresolved $o_a \in O_A$ such that no link was assigned in Pass 1, define the similarity-eligible candidate set as:

$$C_s(o_a) = \{o_p \in O_P : s(o_a, o_p) \geq t\}.$$

If $|C_s(o_a)| \geq 1$, let $o_p^*(o_a)$ represent the single element from $C_s(o_a)$ with the greatest $s(o_a, o_p)$ and include $(o_a, o_p^*(o_a)) \in L_O$. If $C_s(o_a)$ contains elements that are tied with the greatest $s(o_a, o_p)$, let $o_p^*(o_a)$ resolve to the element with the greatest $g(o_a, o_p)$ and include $(o_a, o_p^*(o_a)) \in L_O$.

L_A – Award Linkage Process

This section describes the formally defined process utilized to derive the L_A mapping. As established, let A be the set of award entities in the USASpending layer and P be the set patents in the USPTO layer.

The complexity of the award linkage process is significantly less than the previously defined organizational linkage process due to the inclusion of the legally mandated government interest statement captured within the USPTO dataset. Since 1981, all new patents associated with a government award must identify the award within the patent application if the contract contains the standard patent rights clauses (Rights to Inventions, 2023). This data field is subject to the same limitations as the rest of the patent application and misspellings, accidental omissions, or other data-entry errors are common. Given awards have alphanumeric identifiers and not names, these inconsistencies cannot be resolved through the same process applied within the L_O pipeline. We discovered that the final temporal constraint was effective in resolving many of these errors. This is a significant limitation of the L_A approach. Figure 4 summarizes the L_A data pipeline.

The filtering stage identifies the primary candidate populations from both data layers. We search the USASpending database for unique *award_id's* with a signed-on date between 2014 and 2025. Simultaneously, we search the USPTO database for unique *patent_id's* designated as having a government interest with an application date between 2014 and2025.

In the screening stage we establish initial *award_id* – *patent_id* pairs. A pair is created where the *contract_award_number* extracted from the patent's government interest metadata and the *award_id* from the USASpending data are an exact match. To account for minor data-entry variations, these fields have been normalized by removing special characters and whitespace prior to matching.

In consideration of the patent award data field errors, we further constrain our set by applying a causal temporal constraint to the screened pairs. A candidate pair is considered eligible only if the *award_id* signed-on date strictly precedes the *patent_id* application date. This ensures that the innovation output (the patent) is logically and chronologically attributable to the preceding government award.

The final phase filters the eligible pairs to isolate those relevant to the SBIR ecosystem. Similar to the L_O mapping, *award_id* – *patent_id* pairs are included in the final dataset if the award carries a SBIR I/II/III designation. The formal logic driving the L_A process is as follows.

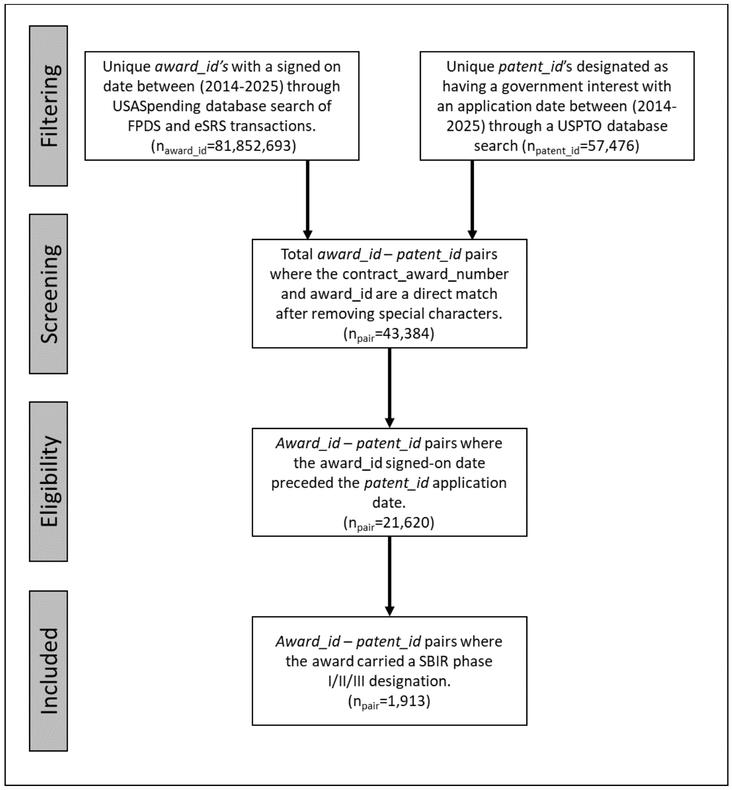

Figure 4: Award linkage process. The LA process utilized a four-stage pipeline to derive cross layer matches between the USASpending and USPTO patent layers. The matching strategy employed is largely deterministic leveraging post-normalized exact string matches.

Unlike the L_O mapping, which is constrained to a many-to-one relationship to resolve organizational hierarchies, the L_A mapping is characterized by full cardinality. Where both patents and awards may carry multiple linkages.

$$\forall a \in A: |\{p \in P: (a,p) \in L_A\}| \; is\ unrestricted$$

$$\forall p \in P: |\{a \in A: (a,p) \in L_A\}| \; is\ unrestricted$$

For each $(a,p) \in A \times P$ let $i_a(a)$ denote the normalized award_id of award a and let $gov_p(p)$ denote the set of normalized PIIDs listed in the government interest statement of patent p. Define the screened pair set:

$$S_A = \{(a,p) \in A \times P: i_a(a) \in gov_p(p)\}$$

To enforce the chronological constraint where the patent must reference a valid award, let $t_a(a)$ denote the signed date of award a and $t_p(p)$ denote the application date of patent p. For each $(a,p) \in S_A$, if $t_a(a) < t_p(p)$, include $(a,p) \in L_A$.

Discussion and Results

The preceding section provided the architectural framework necessary to generate a set of L_A and L_O linkages through which we can now derive sets of SIBR related patents. We begin by constructing three patent sets from the derived mappings (Figure 5). $L_A(p)$ counts patents linked directly to SBIR awards through *award_id – patent_id* pairs and can be interpreted as SBIR award–attributable inventive output. $L_O(p)$ counts patents linked to organizations that won SBIR awards through *uei – assignee_id* pairs and can be interpreted as the broader inventive output of award-winning SBIR organizations. $L_O(p^*)$ is a restricted subset of $L_O(p)$ that retains only patents with a government interest statement as recorded in the Government Patent Register (Gross & Sampat, 2025). $L_O(p^*)$ can be interpreted as an organization-level measure of public-interest inventive output among SBIR award winners.

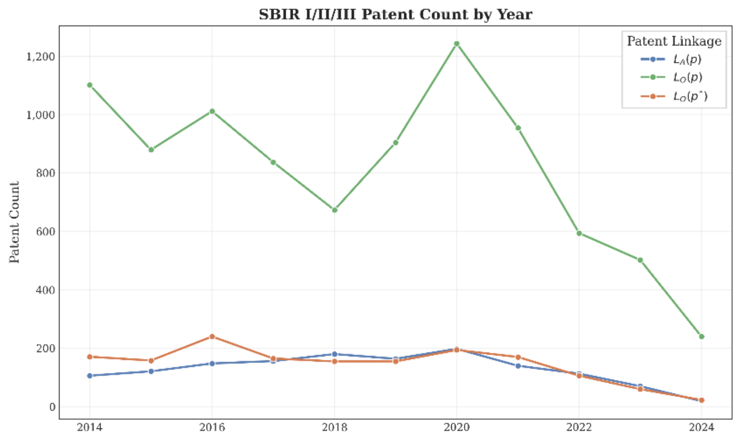

Figure 5: SBIR I/II/III patents by year. Annual patent counts are shown for $L_A(p)$ (award-attributable patents linked directly to SBIR awards), $L_O(p)$ (all patents associated with SBIR award-winning organizations), and $L_O(p^*)$ (the public-interest subset of $L_O(p)$ identified via government interest statements). The magnitude differences highlight the precision–scope trade-off between award- and organization-based linkage, while the close tracking of $L_O(p^*)$ and $L_A(p)$ provides an internal consistency check on the linkage logic. The sharp decline in the most recent years reflects visible patent processing/disclosure lags rather than an abrupt change in underlying inventive activity.

The results of this exercise illustrate the central precision versus scope trade-off implied by the $L_A - L_O$ framework: the chosen linkage function inherently defines the classification boundaries of the downstream patent set, and no single linkage can simultaneously maximize both precision and scope. Figure 5 illustrates this trade-off clearly. $L_A(p)$ yields a comparatively small series because it applies a narrow boundary by explicitly tying patents to specific SBIR awards. $L_O(p)$ produces a much larger series because it shifts the unit of linkage from the award to the organization and therefore includes patents that are not explicitly award-attributable. Additionally, given the size of the SBIR population, $L_O(p)$ should be expected to exhibit higher baseline variability than $L_A(p)$ because it reflects organization-level patenting rather than award-level attribution.

The public–private distinction motivates the role of $L_O(p^*)$. $L_O(p^*)$ tightens the organization-level boundary inherent to $L_O(p)$ by restricting the set to only include public-interest patents. While $L_O(p^*)$ does not provide the award-level precision of $L_A(p)$, it reduces the inclusion of purely private inventive activity while preserving an organization-level perspective. In this sense, $L_O(p^*)$ provides a policy-relevant bridge between the two linkage strategies by retaining some of the breadth of L_O while reducing some of the noise introduced by private patenting that $L_O(p)$ necessarily admits.

This leads us to the second key result of this analysis. For SBIR award-winning organizations, $L_O(p^*)$ tracks $L_A(p)$ closely over time. This alignment functions as an internal validation check on the $L_A - L_O$ logic: two measures derived from different units, awards for L_A and organizations for L_O, produce similar trends in this population. Interpreted through the framework that we have established, this similarity indicates that for SBIR award winners, the subset of patents that can be indicative of public-interest inventive outputs closely corresponds to the inventive activity that can be directly attributed to SBIR awards. Because the SBIR program targets small businesses and nontraditional entrants, it is plausible to expect the government-interest portion of these firms' patenting activity to overlap with the same patenting activity captured by direct SBIR award attribution. This agreement between $L_O(p^*)$ and $L_A(p)$ therefore highlights the complimentary nature of the $L_A - L_O$ framework and, in this constrained setting, the results support the internal consistency of the linkage boundary logic.

While the data presented in Figure 5 are useful for illustrating the magnitude of the framework's precision-scope trade-off in raw counts, direct comparison of the series is difficult because each is defined by a different classification boundary. To provide a complementary view that is more stable to scale differences and better suited for cross-year comparison, we normalized each series by an appropriate "total patent" denominator. This produces comparable ratio measures that retain the conceptual distinctions of $L_A(p)$, $L_O(p)$, and $L_O(p^*)$ by expressing each as a share of its relevant patent universe. This normalization also helps mitigate some of the late-year downward bias from patent processing and disclosure lags by anchoring the numerator to a denominator subject to similar timing dynamics.

Figure 6 reports these normalized series presenting each linkage-based patent set as a proportion of total patents in its corresponding universe. The results show that patenting associated with the SBIR program has generally trended upward in relative terms over the past decade regardless of linkage strategy. This outcome reinforces the $L_A - L_O$ framework by demonstrating how the complementary functions can be jointly employed to view the same phenomena through different boundary definitions. In this case, the agreement across measures lends robustness to the suggestion that the SBIR-linked inventive activity constitutes a growing share of the government-interest patent record.

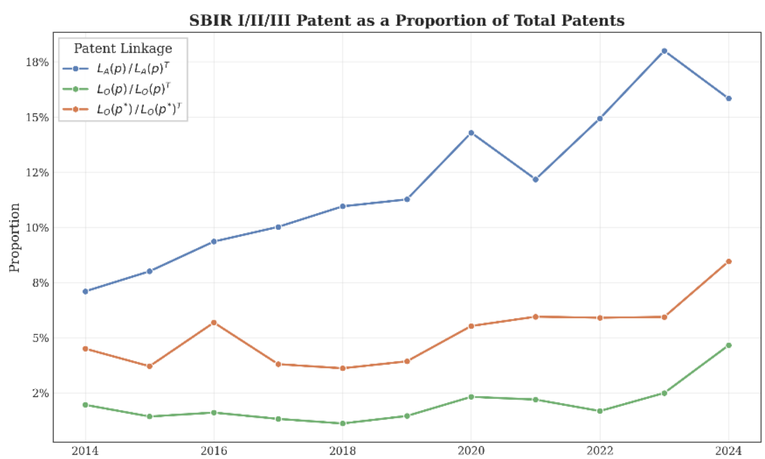

Figure 6: SBIR I/II/III patents as a share of total patents. Each series expresses the corresponding linkage-based patent set as a proportion of its relevant patent universe: $L_A(p)/L_A(p)^T$ for award-attributable patents, $L_O(p)/L_O(p)^T$ for organization-linked patents among SBIR award winners, and $L_O(p^*)/L_O(p^*)^T$ for the public-interest subset identified via government interest statements. Normalizing by totals preserves the classification distinctions while enabling direct comparison of trajectories. Across all three measures, SBIR-associated patenting trends upward as a share of the broader patent record.

Conclusion

Explicitly identifying the trade-offs between award linkage L_A and organization linkage L_O is not merely an academic exercise; it is directly relevant to current U.S. defense industrial policy. As the DoD expands engagement with nontraditional contractors, small businesses, and commercial innovation hubs, interpreting innovation signals becomes increasingly important to policy success. The Secretary of War's (2026) *Transforming the Defense Innovation Ecosystem to Accelerate Warfighting Advantage* memorandum reinforces this orientation by emphasizing data-informed decision-making and accountability for measurable outcomes. Herein lies the fundamental challenge this paper addresses: if inventive signals are used to inform policy decisions, the linkage methodology used to map outcomes must be explicit and well understood.

The organizations targeted by modern initiatives sit precisely at the fault line between the linkage strategies we have defined. Without clarity on the linkage boundary, policy-makers risk drawing flawed conclusions from empirical results. When research is anchored in award linkage (L_A), it can provide high-precision insights into the outcomes that can be directly attributed to specific awards or initiatives. At the same time, policy-makers should be hesitant to extrapolate L_A-based findings into forward-leaning conclusions about system-wide innovation trends or the overall inventive capacity of participating firms. Because L_A captures only those patents that explicitly reference government identifiers, broader inventive activity and spillovers are largely unobserved.

For broader questions relating to organizational capacity, ecosystem effects, or knowledge spillovers, policy-makers should look to research employing organization linkage (L_O) due to its wider analytical aperture. Contrary to L_A, however, defense-specific conclusions drawn from L_O must be treated with caution because of the structural risk of over-classification. Large firms newly entering the DIB may bring extensive patent portfolios in consumer electronics or unrelated commercial products. For these firms, L_O will inherit private efforts alongside public-interest invention. This breadth can be analytically useful (for example, in identifying innovative potential in dual-use technology firms), but it can obscure the marginal effects if the objective is specific to defense-specific activity.

The SBIR demonstration shows how these boundary choices translate into observable differences and also how the linkages can be used jointly to improve interpretability. In raw counts, $L_O(p)$ yielded a much larger series than $L_A(p)$, consistent with its organization-inheritance boundary, while $L_A(p)$ provided a narrow measure of award-attributable inventive output. Restricting $L_O(p)$ to government-interest patents ($L_O(p^*)$) produced a trajectory that closely tracks $L_A(p)$ providing an internal diagnostic that the boundary logic behaves as expected in a population where public-interest invention should plausibly overlap award-attributable activity. Finally, presenting each measure as a share of an appropriate patent corpus provided just one specific example of how the data derived from the framework could support different analytical perspectives highlighting a key benefit of the framework.

Existing defense innovation literature illustrates that scholars have navigated the structural binds inherent to the discussed data-matching challenge pragmatically, but the field would benefit from explicitly declaring the underlying linkage strategy embedded in its research designs. Existing work already recognizes and mitigates the limitations of each approach; we argue these boundary considerations should be treated as first-order methodological choices and made explicit as standard practice. The contribution of this paper is to provide a formal linkage framework that clarifies the trade-offs between the two dominant strategies so that defense innovation research can be compared on common footing, and policy conclusions can be interpreted through an appropriate lens.

Future work will undoubtedly continue to leverage hybrid strategies and advanced classification methods to mitigate trade-offs and data shortfalls, but the core measurement dilemma is structural and thus likely to remain fundamental. Under current data constraints, no procedural refinement can make a single linkage simultaneously maximize attribution precision and system-level scope. By formalizing this trade-off, demonstrating how the two linkages can be used together both as measures and as diagnostics, and by providing an architectural example of a reproducible data pipeline, this paper aims to strengthen the rigor, comparability, and policy relevance of defense innovation analysis grounded in patent–procurement linkages.

References

Bottai, C., de Rassenfosse, G., & Raiteri, E. (2025). A new approach to measuring invention commercialization: An application to the SBIR program. *Research Policy*, *54*(9), 105302. https://doi.org/10.1016/j.respol.2025.105302

Bruce, J., de Figueiredo, J., & Silverman, B. (2018). *Public contracting for private innovation: Government expertise, decision rights, and performance outcomes*. National Bureau of Economic Research. https://doi.org/10.3386/w24724

Decarolis, F., de Rassenfosse, G., Giuffrida, L. M., Iossa, E., Mollisi, V., Raiteri, E., & Spagnolo, G. (2021). Buyers' role in innovation procurement: Evidence from US military R&D contracts. *Journal of Economics and Management Strategy*, *30*(4), 697–720. https://doi.org/10.1111/jems.12430

Defense Contract Audit Agency. (2019). *Cost guidebook*.

De Rassenfosse, G., Jaffe, A., & Raiteri, E. (2019). The procurement of innovation by the U.S. government. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0218927

Fischer, L. F. (2023). The missing link(s) government contracting and patenting. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4347856

Giuffrida, L. M., & Raiteri, E. (2023). *Bureaucratic frictions and innovation procurement*. RePEc. www.RePEc.org

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, *28*(4), 1661–1707. https://doi.org/10.3386/w3301

Gross, D. P., & Sampat, B. N. (2025). The government patent register: A new resource for measuring U.S. government-funded patenting. *Research Policy*, *54*(1), 105142. https://doi.org/10.1016/j.respol.2024.105142

Hall, B., Jaffe, A., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. National Bureau of Economic Research. https://doi.org/10.3386/w8498

Hart, O., Shleifer, A., & Vishny, R. W. (1997). The proper scope of government: Theory and an application to prisons. *The Quarterly Journal of Economics*, *112*(4), 1127–1161. https://doi.org/10.1162/003355300555448

Hicks, K. (2023). *National defense industrial strategy*.

Howell, S., Rathje, J., Van Reenen, J., & Wong, J. (2021). *Opening up military innovation: Causal effects of reforms to U.S. defense research*. National Bureau of Economic Research. https://doi.org/10.3386/w28700

Howell, S. T. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review*, *107*(4), 1136–1164. https://doi.org/10.1257/aer.20150808

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. The MIT Press. https://doi.org/10.7551/mitpress/5263.001.0001

Monath, N., Madhavan, S., Dipietro, C., Mccallum, A., & Jones, C. (n.d.). *Disambiguating patent inventors, assignees, and their locations in PatentsView*. Retrieved from https://bulkdata.uspto.gov/

OECD/Eurostat. (2018). *Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation* (4th ed.). OECD Publishing. https://doi.org/10.1787/9789264304604-en

Raiteri, E. (2018). A time to nourish? Evaluating the impact of public procurement on technological generality through patent data. *Research Policy*, *47*(5), 936–952. https://doi.org/10.1016/j.respol.2018.02.017

Rascona, P., Young, C., & Clowers, N. (2023). *Federal spending transparency: Opportunities to improve USAspending.gov data* (GAO-24-106214). GAO.

Rights to Inventions Made by Nonprofit Organizations and Small Business Firms Under Government Grants, Contracts, and Cooperative Agreements, Patents, Trademarks, and Copyrights, 37 C.F.R. 401 (2023).

Secretary of War. (2026). *Transforming the defense innovation ecosystem to accelerate warfighting advantage*.

Acquisition Research Program
Department of Acquisition, Finance, and Manpower
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net