How many documentary sources sustain Nvidia's lead in reported AI investments?

Lluis Eriksson

19 November 2026

Abstract

Counting the distinct companies in which an investor is reported to have invested puts Nvidia first in a recent public AI supply-chain dataset. We ask how many documentary sources must be omitted before that ranking can change. We retain direction, count a counterparty once, and retain a relationship as long as any supporting source survives. With case-normalized actor names, any 37 hostnames can be omitted without displacing Nvidia. There is a set of 38 whose omission leaves Nvidia and Sequoia Capital tied for first with 43 counterparties each. A second computation reconstructed the graph, verified integer certificates covering every rival, and recounted the witness. Preserving capitalization instead yields a threshold of three because two spellings of Nvidia compete. We also test the effect of treating hosts with matching dated headlines as a single omission. The estimand is the resilience of a ranking in a fixed documentary record. It does not measure amounts invested or the completeness of the record.

1 The ranking and the evidence behind it

A report of an investment can appear on an investor's website, a news service, and a financial portal. Counting articles can therefore conflate editorial repetition with economic activity. Counting distinct counterparties addresses one part of the problem. It leaves a second question: how much of the source base supports the resulting ranking?

We use the public data accompanying Cen et al.'s map of the AI supply chain [1, 2]. Their study characterizes dependence and concentration across industry layers, explores alternative definitions of an actor, and documents the coexistence of names such as Nvidia and NVIDIA. Our question is conditional on their record. We hold the reported relationships fixed and remove units of evidence. A unit can support multiple relationships on either side of the ranking, while one relationship can have multiple supporting units. The minimum number of units sufficient to break a unique lead is a worst-case sensitivity measure. It carries no sampling probability.

2 From reports to supported investments

We used the pinned `gsearch_relationship_results.parquet` file [2], verified by SHA-256 before computation. Of 235,675 rows, 22,555 satisfied the following conditions. The document date fell between 1 November 2022 and 31 October 2026 inclusive. Both actors were companies, the stack layers included capital, and the relationship was an explicit, established investment with a stated direction. We oriented each record from investor to investee. Usable links had an HTTP or HTTPS scheme and a hostname. We excluded empty actor names and loops after normalization.

We applied Unicode NFKC, case folding, and whitespace normalization to names. This resolves capitalization variants but does not identify a parent company with every subsidiary or brand. We lowercased hostnames and removed an initial `www.`. Distinct subdomains otherwise remain separate. The resulting graph contains 6,417 directed pairs, 2,584 distinct investor names, and 707 hosts. The degree of Nvidia is 154. The runner-up, Andreessen Horowitz, has 77 counterparties and Sequoia Capital has 74.

Let H_{ab} be the nonempty set of hosts supporting the directed pair (a, b). For a set S of omitted hosts define

$$d_a(S) = \sum_{b:H_{ab}\neq\varnothing} \mathbf{1}\{H_{ab} \not\subseteq S\}.$$

For initial unique leader L the statistic is

$$k^\star = \min\{|S| : \exists a \neq L, \quad d_a(S) \geq d_L(S)\}.$$

A tie suffices. Removing a host used by both investors may lower both degrees. A large baseline lead by itself gives little information about $k^\star$.

3 A finite certificate for the lower bound

Fix a rival a. It suffices to search over the union of the leader's supports. Restoring an omitted host outside that union cannot restore a leader edge and may restore rival edges. Within the union, group equal support sets and let c_T be the number of leader edges supported exactly on T minus the corresponding number of rival edges. A rival edge with support extending outside the union always survives and makes no contribution to losses. A tie requires

$$\sum_T c_T \mathbf{1}\{T \subseteq S\} \;\geq\; d_L(\varnothing) - d_a(\varnothing).$$

This expression keeps shared evidence on both sides of the inequality.

For each host h let x_h denote its omission indicator. Singletons enter directly. For each other support T, let $0 \le y_T \le 1$. For $c_T > 0$ impose $y_T \le x_h$ for every $h \in T$. For $c_T < 0$ impose $y_T \ge \sum_{h \in T} x_h - |T| + 1$. With integral x the largest attainable left-hand side of the tie inequality is exactly the one obtained by recounting losses. The relaxation $0 \le x \le 1$ admits every real omission. A lower bound on its minimum cost therefore excludes smaller real omissions.

We verified the bounds with integers. Write the rows as $Az \le b$, the objective as $c^\top z$, and the variable box as B. With $D = 10^6$ and any nonnegative integer vector p, every feasible point satisfies

$$Dc^\top z \;\ge\; \min_{v \in B} (Dc + A^\top p)^\top v - p^\top b.$$

Each term on the right is minimized at a box endpoint determined by its sign. In the archived implementation a slack variable on the tie constraint has penalty 1000. A genuine tie uses zero slack. Thus an integer bound strictly above $37D$ excludes every tie costing at most 37. The numerical solver suggests p but its optimality flag is unnecessary for this exclusion.

We tested the ten rivals with degree above 40 individually. To test all other 2,573 together, we granted them the maximum baseline degree in that group, 37, and assumed that all their edges survive. Eleven certificates suffice, each on the original variable box. A separate run rebuilt the graph from the pinned data, checked the integer inequalities without an optimizer, and evaluated the 38-host witness against all investors. That verifier shares the data filter with the discovery code, leaving upstream extraction errors outside the check.

4 What changes the first-place ranking?

The witness leaves Nvidia and Sequoia Capital with 43 documented counterparties each. Every other name has at most 42. The certificates and witness prove $k^\star = 38$. The baseline runner-up need not be the first investor able to catch the leader. The location and redundancy of documentary support matter as well as the initial degree.

If we instead preserve capitalization and trim only outer whitespace, three host omissions suffice for the name `NVIDIA` to overtake `Nvidia`, with 72 and 68 counterparties. An exhaustive enumeration of all 249,571 two-host subsets excluded smaller changes together with the single-host checks. This apparent reversal is a name-resolution artifact. The two different thresholds should be interpreted with their respective definitions of an actor.

5 Co-appearing headlines and limits of inference

We coupled omissions of hosts whose records list the same headline on the same date for the same directed pair. We applied Unicode normalization and case folding, extracted alphanumeric word tokens, and required at least eight words. Repetitions of the same title and date count once for each host pair. Hosts connected by qualifying pairs form a single omission unit, including transitive links. We tested two cutoffs fixed before calculation. The rule measures headline co-appearance and does not establish editorial ownership or identity of the full articles.

Matching rule	Units	Largest unit	Minimum omitted	Tie degrees
Individual hosts	707	1	38	43, 43
At least two title–dates	701	7	35	43, 43
At least one title–date	678	23	32	42, 42

Six host pairs matched at least twice and 31 matched at least once. In each grouped graph the integer lower bound matched a recounted witness, with Nvidia and Sequoia Capital jointly first. The drop from 38 to 32 concerns the number of omission *units*. A unit can contain several hostnames, and relaxing the matching rule can connect hosts that never shared a headline directly. This sensitivity result illustrates how the definition of a source changes the robustness claim.

Publication errors, undated disclosures, private investments, newly added documents, and incorrect investment direction can change the ranking in ways this omission model does not cover. The graph dates records by publication rather than by transaction close. Multiple domains may belong to one owner. Conversely, a single host can carry many separately reported stories. A documentary threshold should consequently travel with its sampling frame, actor definition, and unit of omission.

Studies of centrality under missing data provide a related frame of comparison [3]. Our difference is the observation model in which an omission acts on documentary supports rather than directly on a single network edge. The integer certificate is an application of elementary linear programming duality. The contribution claimed here is the verified threshold and its sensitivity within this fixed dataset.

Model assessments and review context. OpenAI Codex assisted with analysis, software, and drafting in the author's working context. The separate verification run checked a mathematical certificate and recounted the witness. Scientific assessments of this exact PDF, if issued, belong to its hash-bound publication record. No external peer-review assessment is claimed.

Data and reproducibility. The data remain under the upstream CC BY-NC-SA 4.0 license. The accompanying materials provide the pinned input, filters, certificate generator, solver-free checker, witness and grouping analysis. The upstream Parquet file is retrieved from its publisher. Its SHA-256 is given in the scripts.

References

[1] S. H. Cen, L. A. Gailmard, J. Kim, R. Bommasani, D. E. Ho and P. Liang. Mapping the AI supply chain: dependencies, vulnerabilities, and market concentration in the AI ecosystem. arXiv:2611.24927v1.

[2] S. H. Cen and collaborators. AI Supply Chain dataset. Hugging Face, revision e737b94a.

[3] S. P. Borgatti, K. M. Carley and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28 (2006), 124–136. doi:10.1016/j.socnet.2005.05.001.