Entropy and Heterogeneity in the AI Research-to-Product Knowledge Transfer Network

Kranthi Manchikanti · Microsoft

NetSciAI 2026 Satellite

Boston, MA

peergraph.ai

144	91	216	78
Researchers (R)	Products (B)	Adoption edges (E)	Observed domain pairs

A bipartite, builder-declared map of how AI research becomes commercial product · analyzed with Shannon entropy, KL divergence, and mutual information

1 · Network Construction

PeerGraph is a bipartite graph **G = (R ∪ B, E)** with researcher nodes (|R| = 144), product nodes (|B| = 91), and adoption edges (|E| = 216). An edge (r, b) indicates a builder declared product *b* uses a paper by researcher *r*.

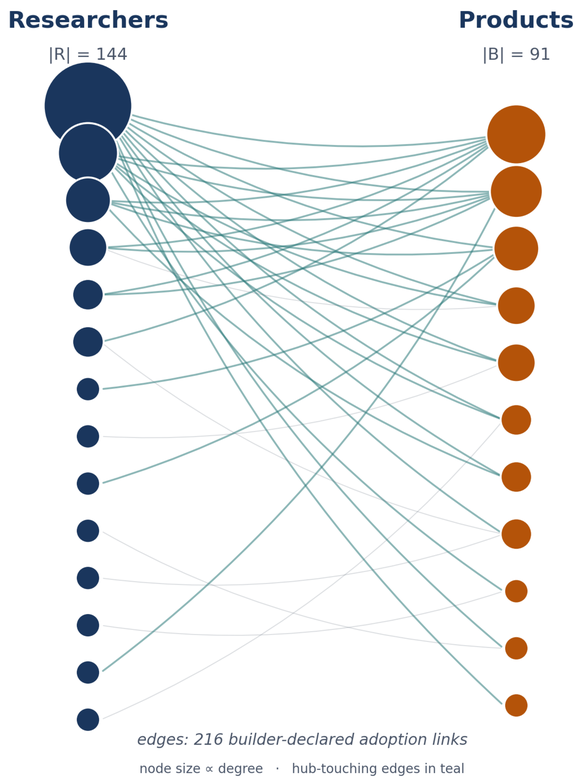
Edges carry domain annotations: 12 research domains × 14 product domains, with 78 of 168 pairs observed at least once.

Degree structure

	Mean	Median	Max	CV
Researchers	1.5	1	58	3.7
Products	2.4	2	11	≈1.6

Researcher CV = **3.7** sits well above the exponential threshold (CV = 1), placing the network in the heavy-tailed regime characteristic of heterogeneous knowledge networks (Newman 2001). Formal distributional fitting (Clauset et al. 2009) requires N > 144.

Schematic



Data provenance

Builder self-reports with public provenance. Convenience sample with salience bias toward prominent papers. Construction parallels patent-to-paper citation networks (Jaffe & Trajtenberg 2002) but captures **declared deployment** rather than inventive activity.

Key observation.
A single researcher accounts for **58 of 216 edges** (27%) on the research side, while the median researcher is cited by exactly one product. This 58×-vs-1 spread between hub and median is the structural fingerprint of the heavy tail and the source of high domain-flow inequality (Gini = 0.72).

Domain coverage.
12 research domains (NLP, CV, ML, RL, GenAI, Agents, Robotics, Theory, Systems, HCI, Security, Sci-ML) × 14 product domains. 78 of 168 pairs are populated; the rest are sparse or empty in the current snapshot.

EDGE SCHEMA

```
{
  "edge_id":        "e_00128",
  "researcher_id":  "r_007",
  "product_id":     "b_044",
  "research_domain": "NLP",
  "product_domain":  "GenAI",
  "paper_doi":      "10.xxxx/example.2024",
  "provenance":     "https://<builder-url>",
  "declared_at":    "2025-08-14",
  "verified":       false
}
```

One row per builder-declared adoption. Domain pairs derived from this schema feed the entropy, KL, and MI calculations in §2.

2 · Information-Theoretic Analysis

We treat the distribution of links over 78 domain pairs as a discrete distribution and compute three measures.

Shannon entropy

$H = -\sum p_i \log_2 p_i = 4.48$ bits

Normalized H / H_{max} = 0.71. Perplexity 2^H = **22.4** — the network concentrates flow through **29%** of available pathways.

KL divergence from uniform

D_{KL}(obs || unif) = 1.80 bits

Mutual information

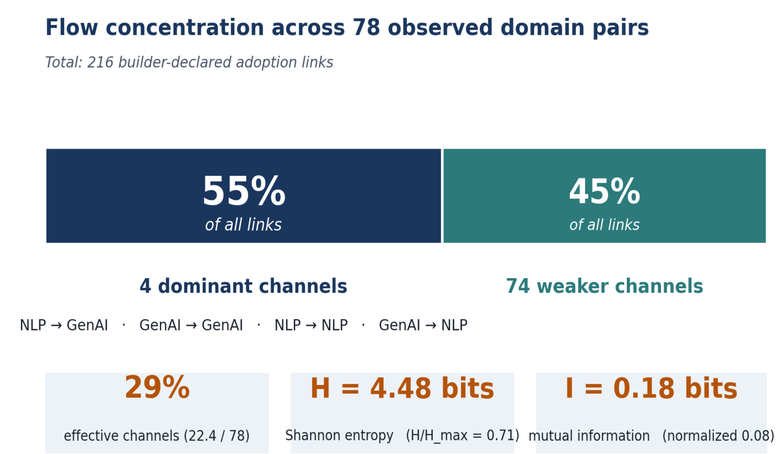
I(src ; tgt) = 0.18 bits (normalized MI = 0.08)

Weak coupling: knowing the research domain provides little predictive information about the product domain. Consistent with the cross-domain transfer ratio of **1.91** (547 cross vs. 287 intra-domain links).

Complexity metrics — summary

Metric	Value	Interpretation
Degree CV (researchers)	3.7	Heavy-tailed regime
Shannon entropy H	4.48 bits (0.71)	Intermediate; structured
Effective channels	22.4 / 78 (29%)	Concentrated pathways
D_{KL}(obs || unif)	1.80 bits	Departure from uniform
I(src ; tgt)	0.18 bits (0.08)	Weak domain coupling
Cross / intra ratio	1.91	Majority cross-domain
Paper adoption Gini	0.57	Moderate-high concentration
Domain flow Gini	0.72	High inequality

Flow concentration



Four channels carry the majority of declared transfer; the remaining 45% is spread across 74 long-tail pairs. The coexistence of dominant channels with a diverse tail is the signature of **structured complexity** — neither random diffusion nor monolithic concentration.

Headline interpretation.
Normalized H / H_{max} = 0.71 places this network between maximal order and maximal disorder — the regime complex-systems theory associates with structured complexity (Feldman & Crutchfield 1998). The system is neither a random graph nor a star: it is a small set of load-bearing channels braided with a long tail.

GET THE DATASET

www.peergraph.ai

License CC0 — no rights reserved
Contents 144 researchers · 91 products · 216 declarations · 150 papers
Schema (researcher, product, $domain_r$, $domain_b$, provenance)
Formats CSV · JSON · GraphML for bipartite tools

scan to access — open & versioned

3 · Higher-Order Structure

Products adopting *k* papers define **(k − 1)-simplices** in the paper co-adoption complex. With mean product degree 2.4, multi-paper co-adoption is common — higher-order knowledge integration events not reducible to pairwise relationships.

Analysis of this simplicial structure (Betti numbers, persistence diagrams) is planned follow-up work.

Concentration

- Paper adoption Gini: **G = 0.57**
- Domain flow Gini: **G = 0.72**
- Most-adopted paper accounts for **26.9%** of all links

Comparable to inequality in citation distributions (Redner 1998); salience bias in declarations may inflate this concentration.

Toward higher-order analysis.
Pairwise edges undercount integration. A product adopting *k* papers across *k* domains is a single higher-order interaction that no pairwise statistic captures. Persistent homology of the co-adoption complex is the natural next instrument.

4 · Limitations

- Convenience sample (N = 216)
- Unverified builder declarations; salience bias
- Entropy and Gini estimates have wide CIs at this scale
- Distributional fitting deferred pending larger N

Future work

- Expanded coverage with verified provenance
- Bipartite null models (Saracco et al. 2015)
- Temporal entropy evolution
- Topological data analysis of the co-adoption complex

Selected references

[1] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379–423, 1948.
[2] T. M. Cover, J. A. Thomas. *Elements of Information Theory*, 2nd ed. Wiley, 2006.
[3] A. B. Jaffe, M. Trajtenberg. *Patents, Citations, and Innovations*. MIT Press, 2002.
[4] M. E. J. Newman. Scientific collaboration networks. *PNAS* 98(2):404–409, 2001.
[5] A.-L. Barabási, R. Albert. Emergence of scaling in random networks. *Science* 286:509–512, 1999.
[6] A. Clauset, C. R. Shalizi, M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.* 51(4):661–703, 2009.
[7] D. P. Feldman, J. P. Crutchfield. Measures of statistical complexity: why? *Phys. Lett. A* 238:244–252, 1998.
[8] S. Redner. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* 4(2):131–134, 1998.
[9] F. Saracco et al. Randomizing bipartite networks: the case of the World Trade Web. *Sci. Rep.* 5:10595, 2015.
[10] F. Battiston et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* 874:1–92, 2020.
[11] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.* 46(2):255–308, 2009.
[12] N. Otter et al. A roadmap for the computation of persistent homology. *EPJ Data Sci.* 6:17, 2017.

METHODS AT A GLANCE

Shannon entropy $H = -\sum_i p_i \log_2 p_i$
KL divergence $D_{KL}(P||Q) = \sum p_i \log_2(p_i / q_i)$
Mutual information $I(X;Y) = \sum p(x,y) \log_2[\, p(x,y) / p(x)p(y) \,]$
Effective channels 2^H (perplexity of the domain-pair distribution)
Degree CV σ_k / <k>, with <k> averaged over researchers in R
Gini coefficient $G = (\sum_{i,j} |x_i - x_j|) / (2\, n^2 \, \langle x \rangle)$

REPRODUCIBILITY

All materials — dataset, analysis notebook, and computed metric outputs — are released under CC0 at **peergraph.ai**. Every value in this poster is reproducible from the released CSV/JSON. Dataset versioned; current snapshot is **v0.3**.

CITE THIS WORK

```
@misc{peergraph2026,
  title        = {Entropy and Heterogeneity in the
                  AI Research-to-Product Knowledge
                  Transfer Network},
  author       = {Manchikanti, Kranthi},
  year         = {2026},
  howpublished = {NetSciAI 2026 Satellite},
  url          = {https://www.peergraph.ai}
}
```

Open science · CC0 dataset · corrections and submissions welcome at peergraph.ai.

v0.4 ROADMAP

- Grow beyond 500 builder-declared adoption edges
- Verified-source flag for at least half of v0.3 declarations
- Temporal slicing using `declared_at` month buckets
- Bipartite null-model baseline (Saracco et al. 2015)
- Persistent homology on the paper co-adoption complex
- Power-law fitting once N supports Clauset 2009 estimators
- Expand to 14 product domains × 12 research domains uniformly
- Author-disambiguated researcher resolution via ORCID

*Target release Q4 2026 · updates released through **peergraph.ai** · feedback welcome.*

ACKNOWLEDGMENTS

Thanks to the **91 builders** who publicly declared their paper adoptions, without whom this dataset would not exist. To the open-source maintainers behind the scientific Python ecosystem that powers the analysis. To early reviewers for sharpening the framing around heavy tails and mutual information. And to the NetSciAI 2026 program committee.

*Corrections, declarations, or methodology questions welcome at **peergraph.ai**.*