CSC 84040: Advanced Data Mining *with AI Agents*

Lecture 05: *Link Analysis & Web Mining: Citation Network Mining & SciSciNet (PageRank, HITS) in the context of Low-Resource Learning*

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at the City University of New York

Saptarashmi Bandyopadhyay

September 23, 2026 CSC 84040

Today's Lecture: Premise and Goal

Graphs encode human authority, collective attention, and the structure of scientific discovery.

Classical link algorithms assume dense, cyclic topologies with abundant historical signals. Real-world knowledge networks break these assumptions: they form temporally constrained, highly sparse, long-tailed structures where newborn ideas start with zero inbound links.

Tonight's Goal: Bridge classic spectral graph theory (PageRank, HITS) with modern science-of-science data lakes (SciSciNet) and robust, bias-aware neural graph architectures for data-scarce settings.

Saptarashmi Bandyopadhyay

Roadmap & Conceptual Arc

Foundational Link Analysis & Spectral Ranking

01

Taming web-scale anarchy: Random surfers, transition matrices, topological dead ends, and Kleinberg's mutual recursion.

Adversarial Topologies & Robustness

02

The economics of link manipulation: Deconstructing spam farms, mathematical amplification, and trust-propagation filters.

Scientific Knowledge Graphs & SciSciNet

03

Citations as directed acyclic flows: Navigating 134M papers, external funding linkages, and advanced bibliometric metrics (Disruption, Sleeping Beauties, WSB modeling).

The Low-Resource Frontier & Bias Mitigation

04

The newborn paper challenge: Bypassing explicit venue shortcuts using multi-agent semantic extraction and GroupDRO.

Saptarshmi Bandyopadhyay

Foundational Link Analysis & Spectral Ranking

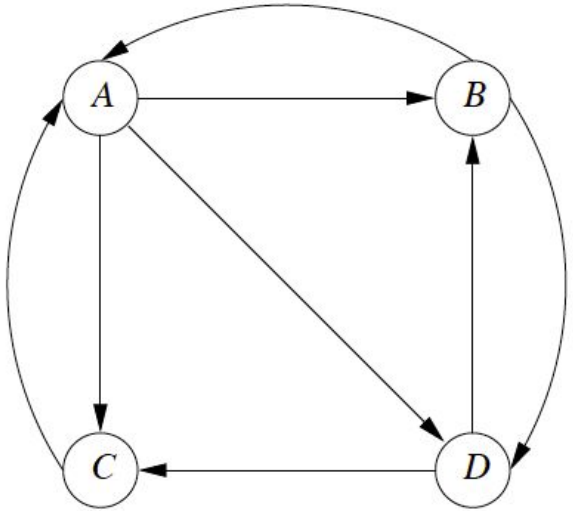
Saptarashmi Bandyopadhyay

Why Counting Links Fails

Early search engines indexed keywords directly from crawled HTML, creating an easy target for "term spam": hidden text, repeated keywords, and stolen content.

Authors link outward to recommend reputable resources. Crucially, you cannot easily forge what other independent webmasters say about your page.

A citation from a major, heavily endorsed hub (like Wikipedia or a major university) conveys orders of magnitude more authority than an endorsement from a newly minted blog.



Mining of Massive Datasets, Chapter 5, Figure 5.1.

A 4-node directed graph (nodes A, B, C, D) illustrating cross-links, cyclic flow, and mutual endorsement among web pages.

The Math of Random Surfers: Power Iteration

Formulating the Walk: Imagine a web surfer clicking random links indefinitely. If a page has k outgoing arcs, each link is traversed with probability 1/k.

The Transition Matrix: Let matrix M represent graph G: entry M_ij = 1/k if page j links to page i, and 0 otherwise. M is column-stochastic (its columns sum to 1).

Convergence to Equilibrium: Under the Perron-Frobenius theorem, if the graph is strongly connected and aperiodic, probability vector v converges to the stationary distribution satisfying the eigenvalue equation: $v = Mv$

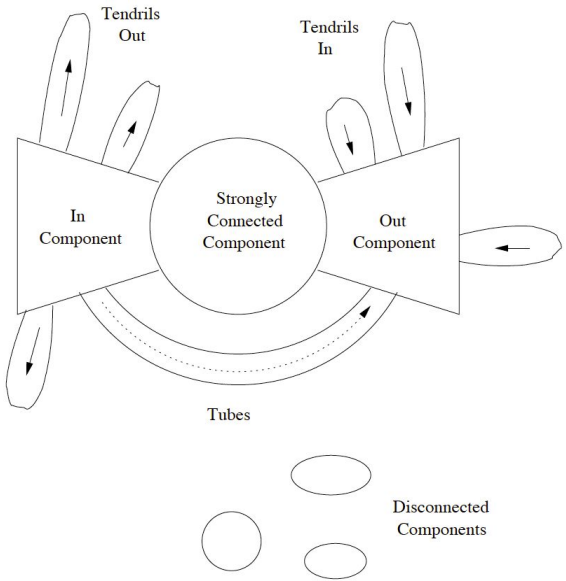
Why Not Gaussian Elimination? Solving (M - I)v = 0 via elimination takes $O(n^3)$ operations. When n is in the billions, iterative sparse matrix-vector multiplication is the only viable path.

Saptarshi Bandyopadhyay

Pathologies of the Real Web

Broder et al.'s classic "bowtie" study revealed the web is fragmented into an SCC, large In/Out components, dangling tendrils, and isolated clusters.

- **Dead Ends Bleed Rank to Zero**
  a. Pages with zero out-links turn M substochastic (column sum = 0). As surfers reach these dead ends, total probability mass drains from the graph with every iteration until all ranks hit zero.
- **Spider Traps Monopolize Attention**
  a. Sinks, self-loops, or tight cliques with no exterior exit suck in random surfers and never let them leave, capturing 100% of global PageRank.



Sahasrakshmi Bandyopadhyay

Fixing Spiders & Dead Ends with Teleportation

At each step, a surfer follows a hyperlinked arc with probability β (commonly 0.80 to 0.85) or teleports to a completely random page with probability 1 - β.

The Regularized Formulation:
Teleportation converts the graph into an irreducible, strongly connected Markov chain:

$$v' = \beta M v + \frac{1-\beta}{n}\mathbf{e}$$

If a surfer falls into a spider trap, they escape in expected $(1 - \beta)^{-1}$ steps. If they hit a dead end, uniform allocation prevents importance from vanishing.

Saptarashmi Bandyopadhyay

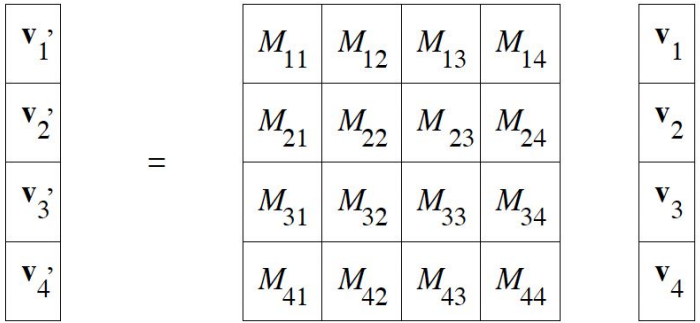
Block-Stripe Matrix Partitioning

When n reaches tens of billions, neither vector v nor transition matrix M fits in memory. Striping M into vertical bands still causes random scattered writes across the entire result vector v', inducing constant page faults.

The Block-Stripe Architecture
Divide matrix M into k × k square blocks, and segment vector v into k stripes.

Memory-Aligned Multiplication
A single Map task processes block M_ij by streaming solely stripe v_j to accumulate into output stripe v'_i:

$$v_i' = \sum_{j=1}^{k} M_{ij} v_j$$



I/O Efficiency
Matrix *M* is streamed from disk exactly once; vector *v* is replicated *k* times across the network, completely avoiding disk thrashing.

Saptarashmi Bandyopadhyay

Personalized Random Walks with Topic-Sensitive PageRank

- **The Search Disambiguation Dilemma:** What does a user mean by searching "Jaguar"? An apex predator, a luxury automobile, or a vintage operating system?

- Instead of teleporting uniformly across all *n* nodes, restrict random restarts strictly to a curated subset of domain pages *S* (e.g., categorized sports or computing directories):

$$v' = \beta M v + \frac{1-\beta}{|S|}\mathbf{e}_S$$

- Surfer probability concentrates around *S* and decays exponentially with graph distance, generating a localized relevance basin.

- **Query-Time Blending:** Compute a compact set of topic vectors offline; at query time, take their linear combination based on the user's recent browsing history.

Saptarashmi Bandyopadhyay

Kleinberg's Alternative: Hubs and Authorities (HITS)

Web resources serve distinct functions: **Authorities** host high-quality content; **Hubs** curate and direct attention toward authorities.

A page is an outstanding hub if it points to authoritative targets; a page is an outstanding authority if trusted hubs link to it:

$$a = L^T h, \quad h = La$$

Substituting one equation into the other yields independent symmetric formulations:

$$a = \lambda\mu(L^T L)a, \quad h = \lambda\mu(LL^T)h$$

Unlike PageRank, HITS handles dead ends and spider traps naturally without artificial dampening.

Saptarashmi Bandyopadhyay

Adversarial Topologies & Robustness

Saptarashmi Bandyopadhyay

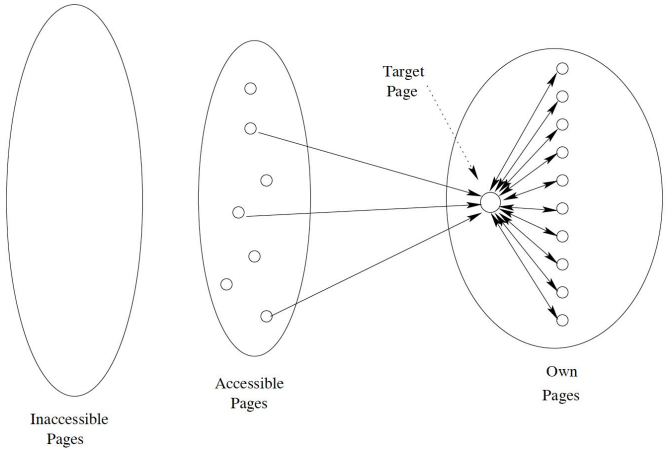
Anatomy of a Link Spam Farm

Once PageRank replaced raw keyword counts, spammers adapted by weaponizing graph topology to funnel artificial prestige to targeted sites.

The Tripartite Spam Blueprint

1. *Inaccessible Web*: High-reputation sites spammers cannot edit.
2. *Accessible Web*: Discussion boards, wikis, and comment fields where spammers paste inbound links.
3. *Owned Pages*: Millions of cheap dummy nodes controlled via automated scripts.

The target page *t* links outward to *m* supporting pages, and all *m* pages link exclusively back to *t*, recycling rank in a closed loop.



Sai Sarashmi Bandyopadhyay

Demonstrating the Spam Exploit

Let target *t* receive external score *x* from accessible blogs, plus rank recycled from *m* supporting pages:

$$y = x + \beta m \left(\frac{\beta y}{m} + \frac{1-\beta}{n} \right) + \frac{1-\beta}{n}$$

Solving for target rank *y* exposes two additive bonuses:

$$y = \frac{x}{1-\beta^2} + c\left(\frac{m}{n}\right), \quad \text{where } c = \frac{\beta}{1+\beta}$$

At standard $\beta = 0.85, 1/(1-\beta^2) \approx 3.60$. The spam topology amplifies outside links by **360%** while harvesting **46%** of its relative web share m/n.

Sudarshan Bandyopadhyay

TrustRank

Spammers can easily link outward to Harvard or the New York Times, but reputable institutions will virtually never link down to a link farm.

We want to establish an elite root set *T* of verified, clean domains (manual editorial reviews, university .edu domains, etc.).

Compute personalized PageRank where the teleport vector is restricted strictly to root set *T*:

$$v_{\text{trust}} = \beta M v_{\text{trust}} + \frac{1-\beta}{|T|}\mathbf{e}_T$$

Trust attenuates by a factor of β on every hop. Because spam farms sit multiple hops away on the graph periphery, their incoming TrustRank approaches zero.

Saptarashmi Bandyopadhyay

Spam Mass Metric to Diagnose Genuine Links

A high PageRank score alone doesn't tell you *how* that rank was acquired. We must contrast ordinary PageRank *r* with verified TrustRank *t*.

Spam Mass quantifies the fraction of a page's importance driven by untrusted topological manipulation:

$$\text{SpamMass}(p) = \frac{r_p - t_p}{r_p}$$

- **Interpreting the Diagnostic Spectrum**
  - Negative / Near Zero: Trust matches or exceeds generic traffic; the node is safe.
  - Close to 1.0: Virtually all rank comes from circular link farms; flag for index expulsion.
- **Production Deployment**
  Modern search engines use Spam Mass as a primary automated heuristic to penalize and de-index search farms at scale.

Sudarshan Bandyopadhyay

Scientific Knowledge Graphs & SciSciNet

Citations as Directed Acyclic Graphs, Multi-Modal Science Lakes, and Impact Dynamics

Saptarashmi Bandyopadhyay

Citations != Hyperlinks

Web pages link freely backward and forward across time, forming cycles. A published paper can only cite historical literature, turning the citation graph into an immutable **Directed Acyclic Graph (DAG)**.

Because new papers enter the graph with zero in-bound citations, naive PageRank systematically assigns all weight to older landmark papers, failing to surface emerging breakthroughs.

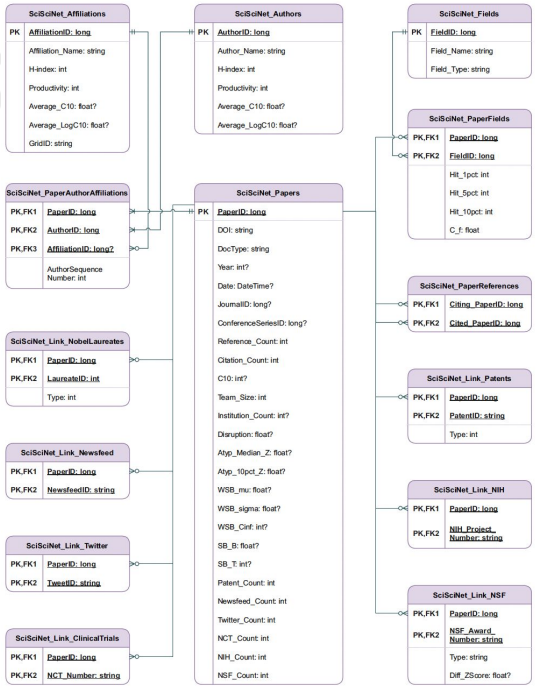
Citation dynamics exhibit extreme preferential attachment: a tiny fraction of foundational papers capture the vast majority of all citations, creating an ultra-skewed distribution.

Saptarashmi Bandyopadhyay

SciSciNet: Architecture of an Open Science Data Lake

A research-ready data lake built on 134,129,188 scientific papers, 134M authors, and 1.58B citation linkages, rescuing and extending the sunset Microsoft Academic Graph (MAG).

- **Upstream Funding Integration:** Directly matches publications to government support: 6.01M links to NIH projects and 1.31M links to NSF awards.
- **Downstream Socioeconomic Tracking:** Traces real-world technology transfer: 38.7M patent citations across USPTO/EPO, 438K clinical trials (NCT), and 55.8M social media engagements.



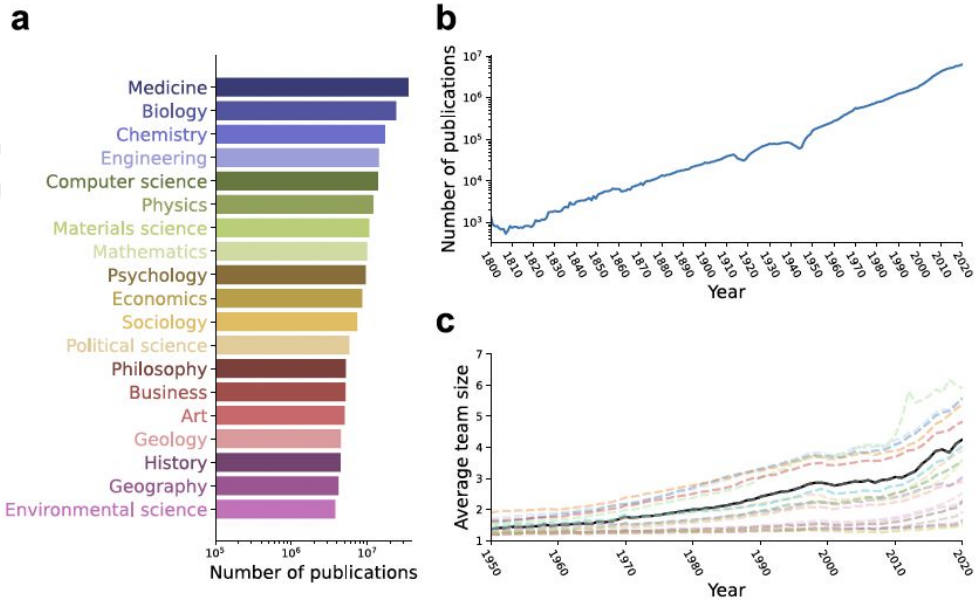
© Lakshmi Bandi Vidhyay

The Macro-Laws of Scientific Knowledge Production

Scientific publication volume has grown steadily since the 1800s across all 19 top-level scientific disciplines.

The average number of authors per paper has climbed from solo investigators in 1950 to large multi-institution teams today.

Societal uses (patents, clinical trial translations, social chatter) follow steep power-law tails spanning multiple orders of magnitude.



The Disruption Index (D)

High citations can mean two completely different things: building incrementally on a paradigm, or destroying that paradigm to establish a new one.

The CD Metric Formalism: Examines how subsequent research cites the focal paper versus its cited foundations:

$$D = \frac{n_i - n_j}{n_i + n_j + n_k}$$

- n_i: Cites *only* the focal paper (disruptive; eclipses past work).
- n_j: Cites *both* focal paper and its ancestors (developmental).
- n_k: Cites ancestors without citing the focal paper.

Empirical Takeaway: Small teams consistently disrupt science (D > 0), while massive teams develop existing frameworks (D < 0).

Saptarashmi Bandyopadhyay

Delayed Recognition

Transformative papers often lie dormant for years (or decades) before experiencing sudden, explosive citation bursts.

The Beauty Coefficient (B): Quantifies a paper's awakening time (t_m) by measuring how far its citation curve bows below a straight line:

$$B = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m} - c_0}{t_m} \cdot t + c_0 - c_t}{\max(1, c_t)}$$

The WSB Generative Model: Predicts lifetime impact c_i^t by parameterizing preferential attachment, aging decay, and fitness.

$$c_i^t = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right)$$

Saptarashmi Bandyopadhyay

The Recombination Theory of Science

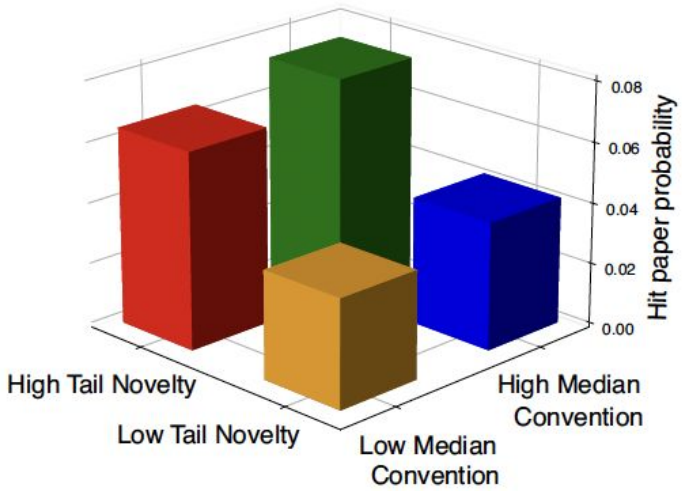
Breakthroughs rarely emerge from complete weirdness or pure routine: they balance familiar discipline foundations with an unexpected twist.

Null Model Co-Citation Shuffling: Construct 10 randomized citation networks that preserve year and field distributions to calculate z-scores for all cited journal pairs.

Median vs. Tail Novelty

- Median Z-Score: Captures how conventional and grounded the paper's foundations are.
- 10th-Percentile Z-Score: Identifies unexpected, atypical combinations of distant fields.

The Winning Quadrant: Papers combining high median conventionality with high tail novelty have a **7.32% probability** of becoming top-5% hit papers.



Saptarashmi Bandyopadhyay

Working with Low Resource Data

Citation Prediction Under Cold-Start Sparsity, Feature Shortcuts, and Group Shifts

Saptarashmi Bandyopadhyay

The Cold-Start Problem: Evaluating Newborn Research

Grant committees and institutions need to evaluate research quality immediately upon publication, long before citations accumulate.

At Day 0, in-degree is strictly zero ($c_0 = 0$). Classic link analysis (PageRank/HITS) and cascade models (DeepCas) fail completely because there is no diffusion path to follow.

Lacking citation histories, models overfit coarse explicit proxies: most notably **venue prestige** (e.g., "accepted at NeurIPS = high impact").

Saptarashmi Bandyopadhyay

Generalization in the Low-Resource Tail

Under active sampling theory, the generalization error disagreement between a candidate hypothesis h_Q and the optimal h* converges with rate:

$$R(h_Q) - R(h^*) \leq \mathcal{O}\left(\sqrt{\frac{d \ln Q + \ln(1/\delta)}{Q}}\right)$$

Label Complexity Upper Bound: Pruning hypotheses using an error disagreement coefficient θ_A bounds required sample size by:

$$N(\mathcal{H}, n, Q, \mathcal{A}) \leq \mathcal{O}\left(\theta_A \left(d \ln Q + \ln \frac{1}{\delta}\right) \ln Q\right)$$

The Problem with Empirical Risk Minimization (ERM)

ERM minimizes average loss across the dataset. In long-tailed citation networks, highly cited papers dominate the loss, causing models to systematically fail on sparse, lowly cited cohorts.

Sanjanashmi Bandyopadhyay

Questions?

Saptarashmi Bandyopadhyay