# SALSA

## Stochastic Approach for Link-Structure Analysis

A comprehensive exploration of fundamental algorithms for web search ranking and link structure analysis

| PR | SALSA | HITS |
|---|---|---|
| PageRank | SALSA | HITS |

**Justin-Marian Popescu | DMDW**

# Introduction and Problem Context

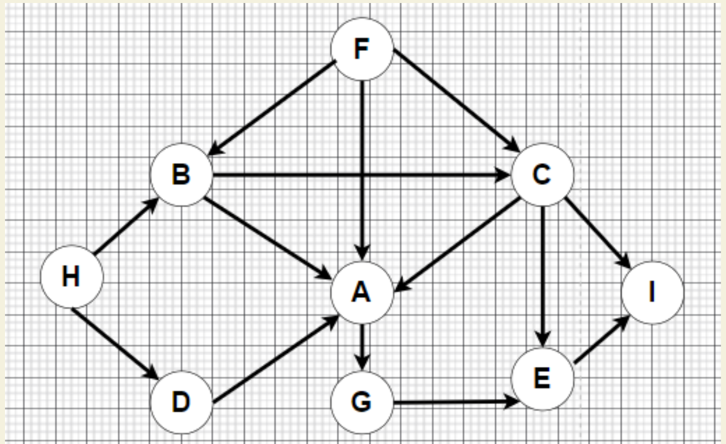
Link analysis algorithms assess the significance of nodes in networks through their connections, offering key methods for web search ranking and structural network analysis.

## Web Search Applications

- **PageRank** formed the foundation of Google's search algorithm
- **SALSA** combines benefits of both approaches for community discovery
- **HITS** provides topic-specific search by distinguishing hubs and authorities

## Graph Theory Applications

- **Centrality measurement** for network analysis
- **Identification of authoritative sources** in information networks
- **Community detection** in social networks




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| B | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| D | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| H | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# PageRank Formulation

## Definition

Given a directed graph with adjacency matrix $A \in \{0,1\}^{n \times n}$, let the row-stochastic transition matrix ***P*** be:

$$P_{ij} = \begin{cases} \frac{1}{d_i^+}, & \text{if } i \to j, \\ 0, & \text{otherwise.} \end{cases}$$

where ***di+*** is the out-degree of node ***i***.

## Damped Random Walk

$$\pi = \alpha P \pi + (1 - \alpha)\nu$$

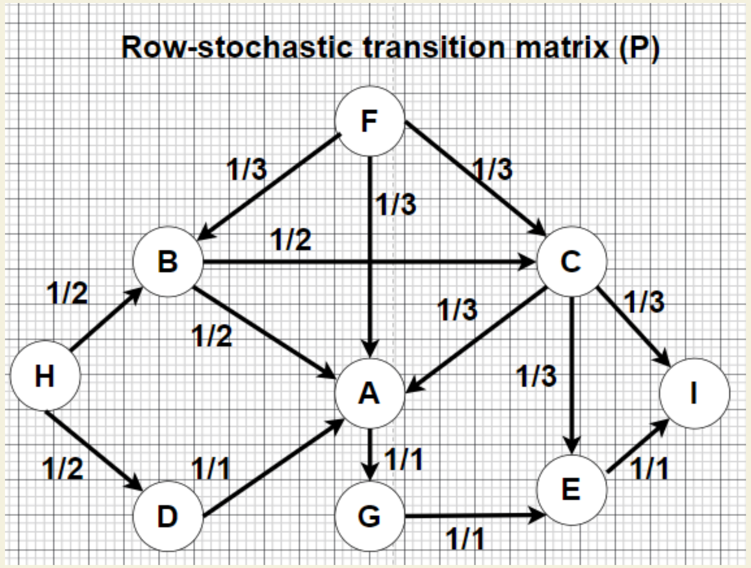
where ***α ∈ (0, 1)*** is the damping factor ***(usually 0.85),***
and $\nu$ (vector) is the teleportation vector (often uniform: ***νi = 1/n***).

## Iterative Computation

$$\mathrm{M} = \alpha P + (1 - \alpha)1\nu^{\top}$$

Convergence is guaranteed because the matrix M is stochastic, irreducible, and aperiodic.

**SALSA (Stochastic Approach for Link-Structure Analysis)**




### Key Properties

- Unique stationary distribution (because of teleportation)
- Captures global influence, not just local structure
- Equivalent to eigenvector centrality on the **Google matrix M**

### Interpretation

***πi*** = steady-state probability that a random surfer visits node ***i.***
A node's PageRank is high if many important nodes link to it, forming a recursive definition of importance.

# PageRank Algorithm Overview

## What is PageRank?

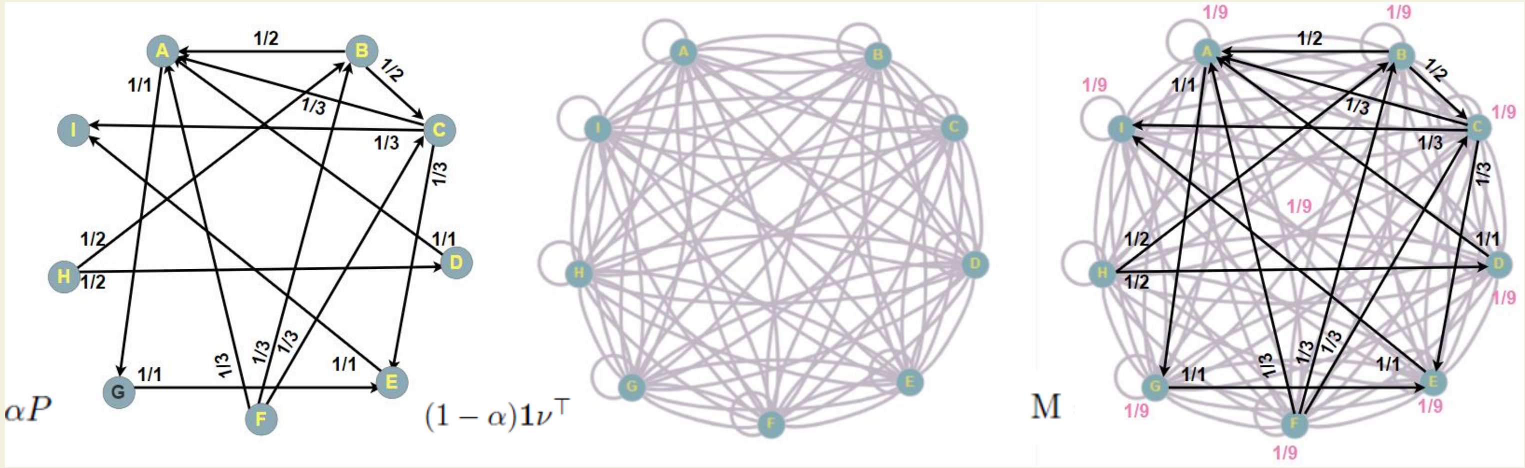
PageRank measures the global importance of web pages based on the probability that a random surfer lands on them.

## Random Surfer Model

- Models a user who randomly follows hyperlinks
- Occasionally jumps to a random page (teleportation)
- Ensures every node can be reached

## Global Importance

- Captures global influence, not just local structure
- Recursive definition of importance
- Forms the basis of Google's search algorithm




### Key Insight

For the added edges must be multiplied with (1- α) and for the existing ones with α.

A node's PageRank is high if many important nodes link to it, forming a recursive definition of importance.

# PageRank Properties and Applications

## Key Properties

### Unique Stationary Distribution
Teleportation ensures a unique steady state probability vector.

### Global Influence Measurement
Captures global importance, not just local neighborhood structure.

### Convergence Guarantees
Matrix $\mathrm{M} = \alpha P + (1 - \alpha)\mathbf{1}\nu^{\top}$ is stochastic, irreducible, and aperiodic.

### Eigenvector Centality
Equivalent to eigenvector centrality on the Google matrix M.

## Practical Applications

### Web Search Ranking
Measures importance of nodes in networks.

### Graph Centrality
Foundation of Google's PageRank algorithm since 1998.

# HITS Algorithm

## Introduction

HITS (Hyperlink-Induced Topic Search) introduces a dual-role link analysis model in which each page is evaluated both as an authority and as a hub.

## Dual Role Model

- **Hub Score:** Good collector of authority pages
- **Authority Score:** Good information source

## Mutual Reinforcement

Hubs point to authorities, and authorities are pointed to by hubs, creating a self-reinforcing cycle.

## Web Context

**Hubs**
Directories or aggregators

**Authorities**
Content-rich pages

### Key Insight

HITS differs from PageRank by using a dual ranking approach. While PageRank assigns a single global importance score, HITS provides two complementary scores that reveal different aspects of a node's role in the network. Developed in 1999.

# HITS Formulation

HITS algorithm computes hub and authority scores by performing eigenvector analysis on the network's adjacency matrix, revealing nodes that best represent information sources and authorities.

## Mathematical Formulation

Authority **scores**: $a = A^\top h$

Hub **scores**: $h = A\ a$

Authority: $a = A^\top A a$

Hub: $h = A\ A^\top h$

## Step-by-step

**1)** Start with initial vectors: $h(0)=1_n$, $a(0)=1_n$

**2)** Repeat until it converges or maximum number of iterations reached:

**2)** Update authority scores: $a(t+1) = A^T h(t)$

**3)** Update hub scores: $h(t+1) = Ah(t+1)$

**4)** Normalize after each step

- Authority of a node: sum of hub scores of nodes linking to it.
- Hub of a node: sum of authority scores of nodes it links to.

### Key Mathematical Properties

- Converges to principal eigenvectors of (hub, authority) $(A\ A^\top, A^\top A)$
- Non-stochastic, may produce multiple disconnected eigenpairs
- Sensitive to local subgraphs and query-dependent subwebs

# HITS Properties and Characteristics

### Non-stochastic Nature

- Unlike PageRank, HITS uses algebraic normalization
- May produce multiple disconnected eigenpairs
- Convergence depends on the graph structure

### Local Structure Sensitivity

- Highly sensitive to local subgraphs
- Query-dependent subwebs can dramatically change results
- May not reflect global network structure

### Dual Ranking System

- Produces 2 separate rankings:
- Hub scores: Quality as a collector of authority pages
- Authority scores: Quality as an information source

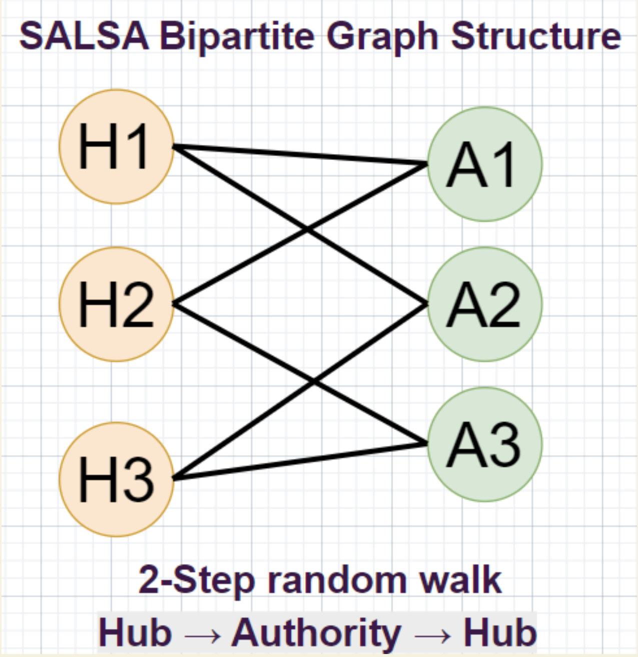
### Key Implications

HITS dual ranking system makes it particularly effective for topic-specific search, identifying both directories/aggregators (hubs) and content-rich pages (authorities). However, its sensitivity to local structures limits its applicability in broader network analysis.

# SALSA Algorithm

## What is SALSA?

SALSA (Stochastic Approach for Link-Structure Analysis) merges PageRank's stability with HITS' role distinction.




Hubs Authorities

## Key Features

- Builds a bipartite graph of hubs and authorities
- Runs two Markov chains on different graph sides
- Uses stochastic normalization for stability
- Performs two-step random walks on the bipartite graph

## Algorithmic Benefits

- Robust against local structure changes
- Detects natural communities in networks
- Produces stochastic hub and authority scores
- Guaranteed convergence

### SALSA vs. PageRank vs. HITS

| PageRank | SALSA | HITS |
|---|---|---|
| Global importance<br>Teleportation | Stochastic roles<br>Bipartite walks | Role separation<br>Eigenvector |

# SALSA Formulation

### Bipartite Graph Construction

Given directed graph G = (V, E):

- *U = {u ∈ V: du > 0} (hubs – offers links to others)*
- *V′ = {v ∈ V: dv > 0} (authorities – receives links from others)*
- Bipartite adjacency (B):

$$(D_U)_{uu} = \sum_{v \in V'} B_{uv} \quad \Rightarrow \quad D_U^{-1} B$$

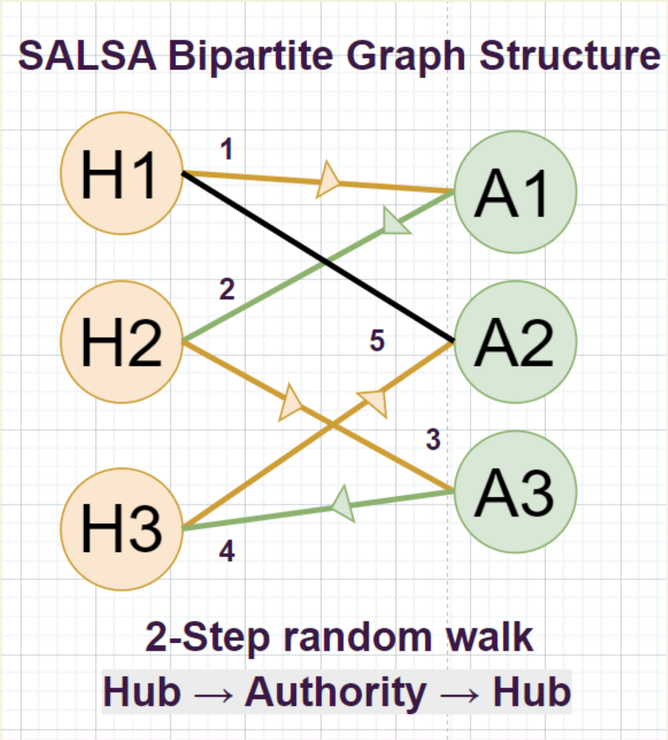
$$(D_V)_{vv} = \sum_{u \in U} B_{uv} \quad \Rightarrow \quad D_V^{-1} B^\top$$

$$B \in \{0,1\}^{|U| \times |V'|}, \qquad B_{uv} = 1 \text{ iff } u \to v,$$

$$D_U = \mathrm{diag}(B\mathbf{1}), \qquad D_V = \mathrm{diag}(B^\top \mathbf{1})$$

$D_U^{-1} B$ normalizes each hub's outgoing links (hub → authority).

$D_V^{-1} B^\top$ normalizes each authority's incoming links (authority → hub).




$$T_{H \to A} = D_U^{-1} B, \qquad T_{A \to H} = D_V^{-1} B^\top$$

Hubs  Authorities

### Two-Step Random Walk

**For hubs:** u → v → w (hub → authority → hub)

**For authorities:** v → u → w (authority → hub → authority)

### Key Properties

- Stochastic normalization ensures convergence
- Robust to local structure changes

# SALSA Properties and Benefits

## Key Properties

### Stochastic Normalization

Random walks with teleportation, ensuring stable results regardless of graph structure.

### Dual Ranking

Produces 2 stochastic role rankings (hubs + authorities) instead of one global ranking.

### Moderate Sensitivity

More robust than HITS but still captures local structure information effectively.

## Key Benefits

**Community Detection**
Identifies natural hub and authority clusters

**Stability**
Robust to graph structure changes

**Role-Based Ranking**
Hubs & authorities as separate dimensions

**Hybrid Approach**
Combines global and local information

**Transition Matrices Pa & Ph + PageRank applied on them**

$$\boxed{P_h = T_{H\to A}\, T_{A\to H} = D_U^{-1} B\, D_V^{-1} B^\top}$$

$$\boxed{P_a = T_{A\to H}\, T_{H\to A} = D_V^{-1} B^\top\, D_U^{-1} B}$$

# Experimental Setup and Datasets

## Framework & Parameters

| Parameter | Value |
|---|---|
| Damping Factor (α) | 0.85 (PageRank) |
| Maximum Iterations | 1000 |
| Convergence Threshold | 1e-6 |

| | |
|---|---|
| Teleportation Vector | Uniform (1/n) |
| Random Restart Rate | 1-α = 0.15 |
| Normalization | Stochastic (sum=1) |

## Evaluation Metrics

### Accuracy

Proportion of correctly ranked nodes

### Computational Time

Execution time and convergence rate

### Scalability

Performance on varying network sizes

## Dataset Characteristics

### Web Graphs ego-Facebook

- Directed links between web pages
- Global structure analysis

### Social Networks soc-sign-bitcoin-otc

- User-follow relationships
- Community detection

### Academic Networks wiki-Vote

- Citation patterns
- Authority detection

# Comparative Results Analysis

Performance comparison across different graph structures and applications:

| Characteristic | PageRank | SALSA | HITS |
|---|---|---|---|
| Normalization | Stochastic (probabilistic) | **Stochastic (probabilistic!)** | Algebraic (eigenvector) |
| Random walks | Global with teleportation | 2-step on bipartite | None |
| Sensitivity | Global | Moderate | Local (query-based) |
| Output | Global importance | Stochastic hubs & authorities | Hubs & authorities |
| Convergence | Guaranteed (ergodic) | Guaranteed (Markov chain) | Depends on structure |

### PageRank Strengths

- Very stable (teleportation)
- Global importance measurement
- Guaranteed convergence

### SALSA Strengths

- Robust stochastic approach
- Community discovery capability
- Hybrid ranking performance

### HITS Strengths

- Topic-specific analysis
- Clear role separation (hubs/auths)
- Identifies link structure patterns

### Application Recommendations

**Web search ranking:** PageRank
**Link analysis:** PageRank

**Community discovery:** SALSA
**Hybrid ranking:** SALSA

**Topic-specific search:** HITS
**Link structure analysis:** HITS

# Algorithm Comparison Summary

| Comparison Aspect | PageRank | SALSA | HITS |
|---|---|---|---|
| **Normalization** | Stochastic | Stochastic | Algebraic |
| **Convergence** | Guaranteed (Ergodic) | Guaranteed (Markov chain) | Depends on structure |
| **Output Type** | Single global rank | Two stochastic ranks | Two role-specific ranks |
| **Sensitivity** | Global | Moderate, robust | Local (query-based) |

### Unique Characteristics

- PageRank: Global importance via teleportation
- HITS: Dual role model with mutual reinforcement
- SALSA: Stochastic normalization ensures stability

### Key Insight

SALSA uniquely combines PageRank's global approach with HITS' role-based separation, creating a robust algorithm that's sensitive to local structure while maintaining global stability.

### Algorithm Selection

- Web search: PageRank (global importance)
- Topic-specific search: HITS (role separation)
- Community discovery: SALSA (role separation + global importance)

**SALSA** (**Stochastic Approach for Link-Structure Analysis**)

# Discussion and Practical Implications

## Global vs Local Ranking Approaches

### Global Approaches

- **PageRank**: Random surfer model → global importance
- Captures global network influence, not just local structure
- Very stable with teleportation guaranteeing convergence

### Local Approaches

- **HITS**: Topic-specific search → role separation
- Sensitive to local subgraphs and query-dependent subwebs
- May have multiple dominant eigenpairs for bipartite graphs

## Algorithm Selection Criteria

| Algorithm | Best Use Case | Key Advantage | Implementation Note |
|---|---|---|---|
| PageRank | Web search ranking, graph centrality | Global importance with teleportation | Parameter **$\alpha$ = 0.85** |
| SALSA | Community discovery, hybrid ranking | Stochastic role-based with communities | Better for query-independent tasks |
| HITS | Topic-specific search, link analysis | Role separation (hubs & authorities) | May need multiple runs for stability |

### Key Insight 1

Algorithm choice depends on whether you need global importance (PageRank) or role-based analysis (HITS/SALSA).

### Key Insight 2

For practical applications, consider stability requirements and convergence guarantees before algorithm selection.

**SALSA** (**Stochastic Approach for Link-Structure Analysis**)

# Conclusions and Future Directions

### PR PageRank

- Global importance measurement
- Very stable with teleportation

Convergence — Guaranteed

### SALSA SALSA

- Combines PR and HITS benefits
- Community detection capability

Convergence — Guaranteed

### HITS HITS

- Role separation (hubs/authorities)
- Mutual reinforcement mechanism

Convergence — Depends on structure

## Recommendations

**Web Search:**
PageRank for global importance, SALSA for topic communities

**Link Analysis:**
HITS for topic-specific hubs/authorities, SALSA for dynamic communities

**Community Detection:**
SALSA for natural communities, HITS for topic clusters

- Temporal link analysis for evolving networks
- Integration with content analysis for semantic ranking
- Scalable implementations for massive graphs
- Robustness to link spam and adversarial attacks

# References and Bibliography

## PageRank


- Page, L. and Brin, S. (1998).
  Anatomy of a large-scale hypertextual Web search engine.
  *Proceedings of the 7th International Conference on World Wide Web (WWW), 539-550.*
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999).
  The PageRank citation ranking: Bringing order to the Web.
  *Technical Report, Stanford University.*


## SALSA


- Lempel, R. and Moran, S. (2000).
  The stochastic approach for link-structure analysis (SALSA) and the role of authorities.
  *Proceedings of the 9th International Conference on World Wide Web (WWW), 552-561.*


## HITS


- Kleinberg, J.M. (1999).
  Authoritative sources in a hyperlinked environment.
  *Journal of the ACM, 46(5), 604-632.*
- Kleinberg, J.M. (1998).
  Hubs, authorities, and communities.
  *ACM Computing Surveys, 30(2), 173-177.*


## Additional Resources

- SNAP (Stanford Network Analysis Project)
  Collections of datasets and tools for link analysis.
  https://snap.stanford.edu/index.html
- UCLA Office of Advanced Research Computing (OARC)
  https://www.youtube.com/watch?v=V_liCwE_ZoI