




# Social Network Analysis

---

- Social Networks: An Introduction
- Primitives for Network Analysis
- Different Network Distributions 
- Models of Social Network Generation
- Mining on Social Network 
- Summary

# What is New for Link Mining Here

---

- Traditional machine learning and data mining approaches assume:
  - A random sample of homogeneous objects from single relation
- Real world data sets: 
  - Multi-relational, heterogeneous, and semi-structured
- *Link mining / Social network analysis*
  - Newly emerging research area
  - At the intersection of research in social network and link analysis, hypertext and web mining, graph mining, and relational learning

# Information on the Social Network

---

- Heterogeneous, multi-relational data represented as a graph or network
  - Nodes are objects
    - May have different kinds of objects
    - Objects have attributes
    - Objects may have labels or classes
  - Edges are links
    - May have different kinds of links
    - Links may have attributes
    - Links may be directed, are not required to be binary
- Links represent relationships and interactions between objects - rich content for mining

# Metrics (Measures) in Social Network Analysis

---

- **Betweenness:** The extent to which a node lies between other nodes in the network. This measure takes into account the connectivity of the node's neighbors, giving a higher value for nodes which bridge clusters. The measure reflects the number of people who a person is connecting indirectly through their direct links
- **Bridge:** An edge is a bridge if deleting it would cause its endpoints to lie in different components of a graph.

# Metrics (Measures) in Social Network Analysis

---

- **Centrality:** This measure gives a rough indication of the social power of a node based on how well they "connect" the network. "Betweenness", "Closeness", and "Degree" are all measures of centrality.
- **Centralization:** The difference between the number of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the number of links each node possesses.

# Metrics (Measures) in Social Network Analysis

---

- **Closeness:** The degree an individual is near all other individuals in a network (directly or indirectly). It reflects the ability to access information through the "grapevine" of network members. Thus, closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network
- **Clustering coefficient:** A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.


# Metrics (Measures) in Social Network Analysis

---

- **Cohesion:** The degree to which actors are connected directly to each other by *cohesive* bonds. Groups are identified as '*cliques*' if every individual is directly tied to every other individual, 'social circles' if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.
- **Degree (or geodesic distance):** The count of the number of ties to other actors in the network.

# Metrics (Measures) in Social Network Analysis

---

- **Density:** (Individual-level) The degree a respondent's ties know one another/ proportion of ties among an individual's nominees. Network or global-level density is the proportion of ties in a network relative to the total number possible (sparse versus dense networks).
- **Flow betweenness centrality:** The degree that a node contributes to sum of maximum flow between all pairs of nodes (not that node).



# Metrics (Measures) in Social Network Analysis

---

- **Local Bridge:** An edge is a local bridge if its endpoints share no common neighbors. Unlike a bridge, a local bridge is contained in a cycle.
- **Path Length:** The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.

# Metrics (Measures) in Social Network Analysis

---

- **Prestige:** In a directed graph prestige is the term used to describe a node's centrality. "Degree Prestige", "Proximity Prestige", and "Status Prestige" are all measures of Prestige.
- **Radiality:** The degree an individual's network reaches out into the network and provides novel information and influence.
- **Reach:** The degree any member of a network can reach other members of the network.

# Metrics (Measures) in Social Network Analysis

---

- **Structural cohesion:** The minimum number of members who, if removed from a group, would disconnect the group
- **Structural equivalence:** Refers to the extent to which nodes have a common set of linkages to other nodes in the system. The nodes don't need to have any ties to each other to be structurally equivalent.
- **Structural hole:** Static holes that can be strategically filled by connecting one or more links to link together other points. Linked to ideas of *social capital*: if you link to two people who are not linked you can control their communication

# A Taxonomy of Common Link Mining Tasks

---

- Object-Related Tasks
  - Link-based object ranking
  - Link-based object classification
  - Object clustering (group detection)
  - Object identification (entity resolution)
- Link-Related Tasks
  - Link prediction
- Graph-Related Tasks
  - Subgraph discovery
  - Graph classification
  - Generative model for graphs

# What Is a Link in Link Mining?

---

- Link: relationship among data
- Two kinds of linked networks
  - homogeneous vs. heterogeneous
- Homogeneous networks
  - Single object type and single link type
  - Single model social networks (e.g., friends)
  - WWW: a collection of linked Web pages
- Heterogeneous networks
  - Multiple object types and link types
  - Medical network: patients, doctors, disease, contacts, treatments
  - Bibliographic network: publications, authors, venues

# Link-Based Object Ranking (LBR)

---

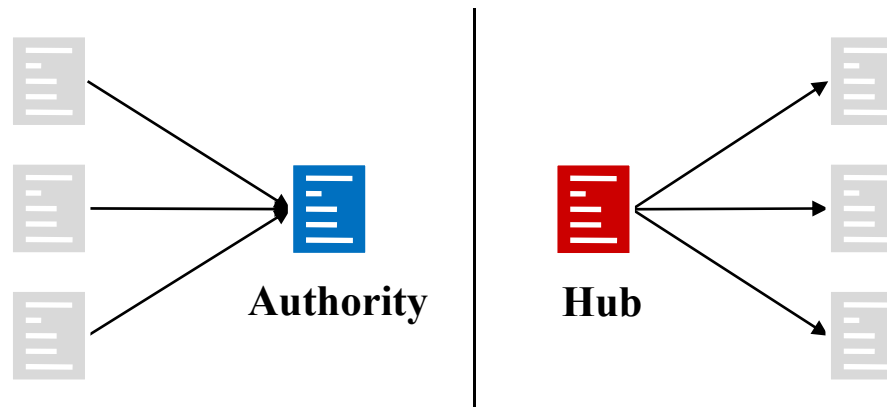
- LBR: Exploit the link structure of a graph to order or prioritize a set of objects within the graph
  - Focused on graphs with single object type and single link type
- This is a primary focus of link analysis community
- Web information analysis
  - *HITS* and *PageRank* are typical LBR approaches



# HITS: Capturing Authorities & Hubs (Kleinberg'98)

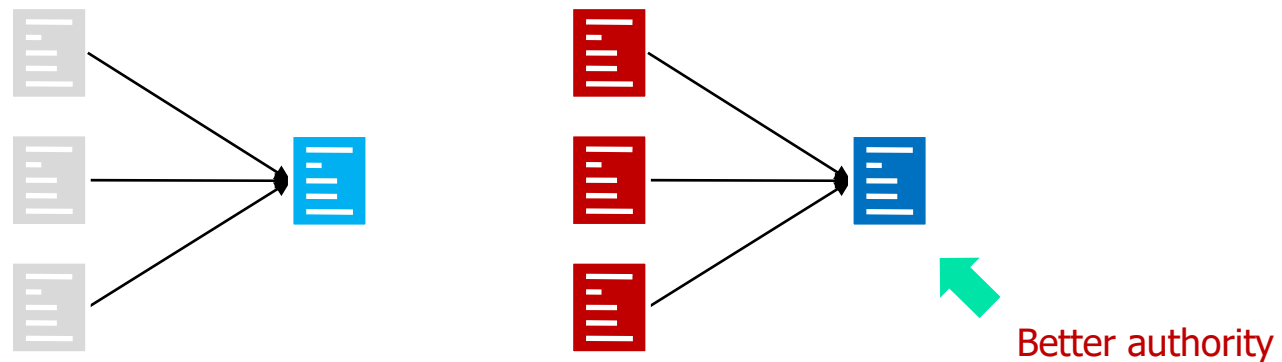
## ■ Intuitions

- Links are like citations in literature
- Pages that are widely cited are good *authorities*
- Pages that cite many other pages are good *hubs*



# HITS: Capturing Authorities & Hubs (Kleinberg'98)

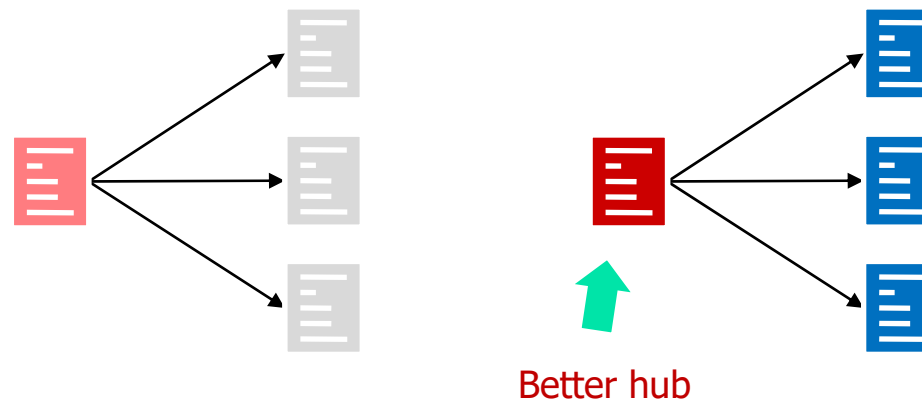
- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - *Iterative mutual reinforcement* ...





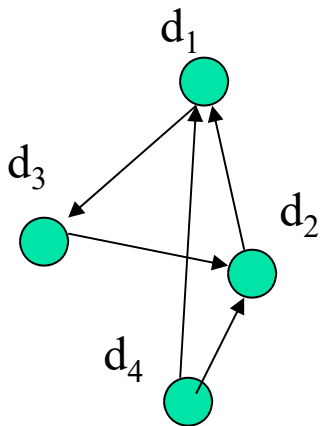
# HITS: Capturing Authorities & Hubs (Kleinberg'98)

- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - *Iterative mutual reinforcement* ...



# The HITS Algorithm (Kleinberg 98)

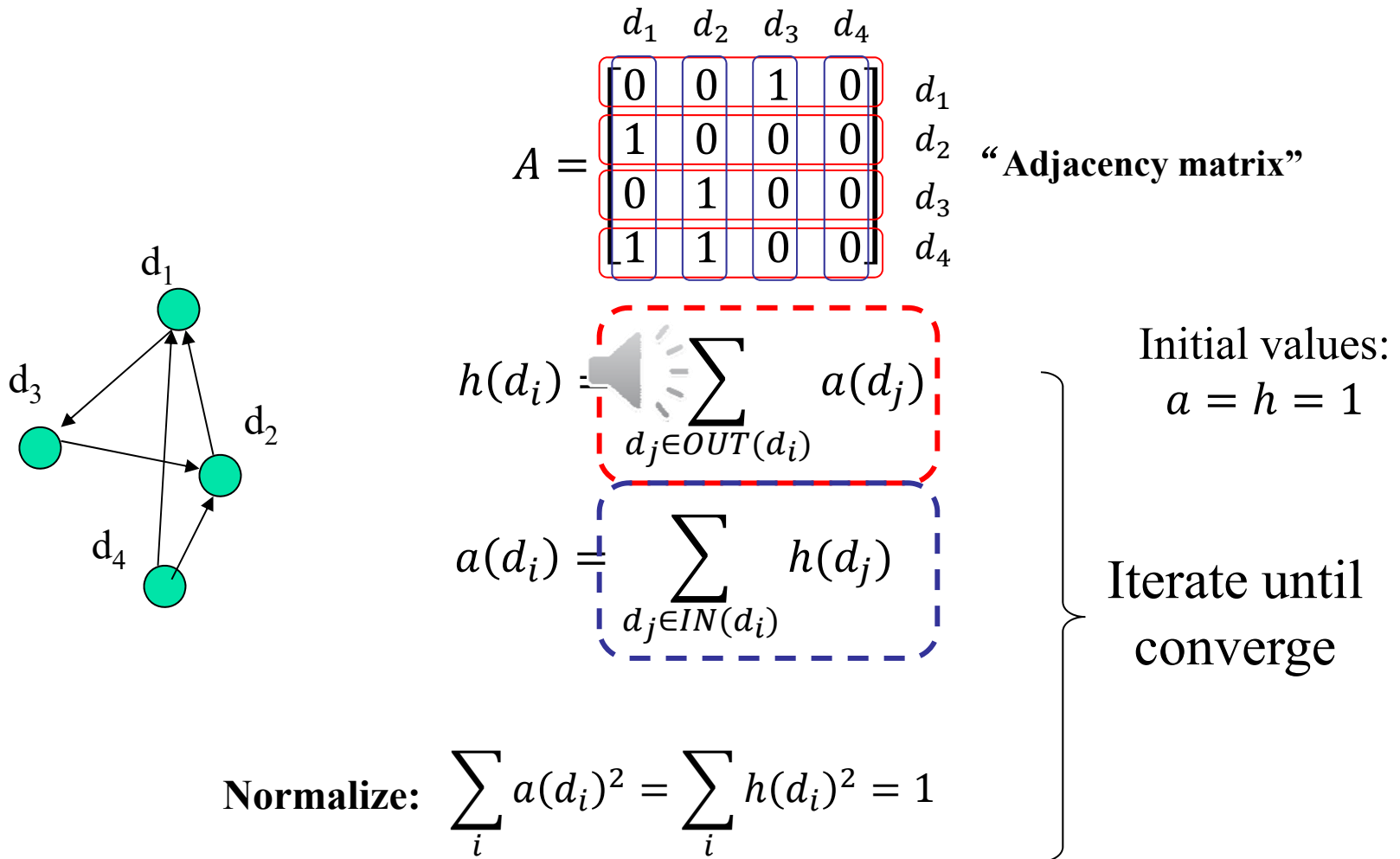
- Each page ( $d_i$ ) has two scores:
  - Hub score ( $h(d_i)$ ) and authority score ( $a(d_i)$ )
  - *Hub score is the sum of authority scores* from its out-link neighbors
  - *Authority score is the sum of hub scores* from its in-link neighbors



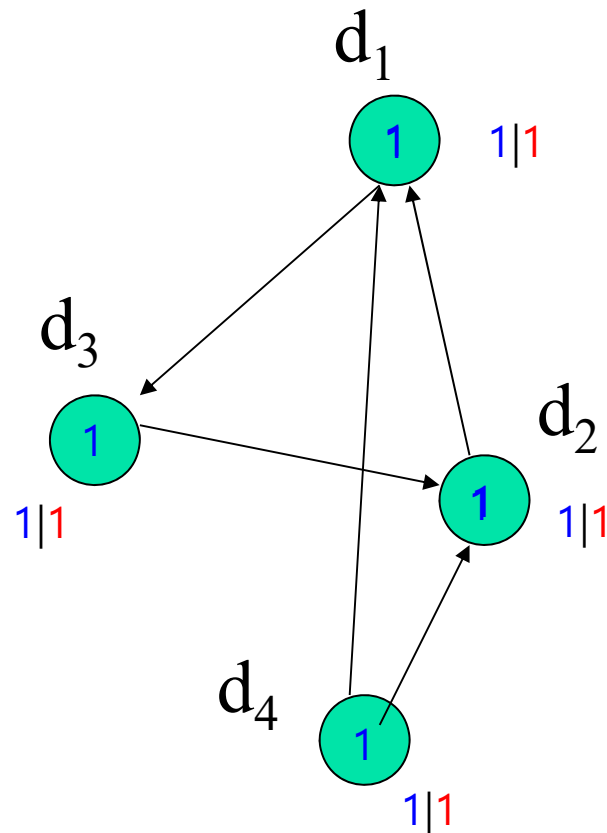
$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j) \quad \text{“Hub score”}$$

$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j) \quad \text{“Authority score”}$$

# The HITS Algorithm (Kleinberg 98)

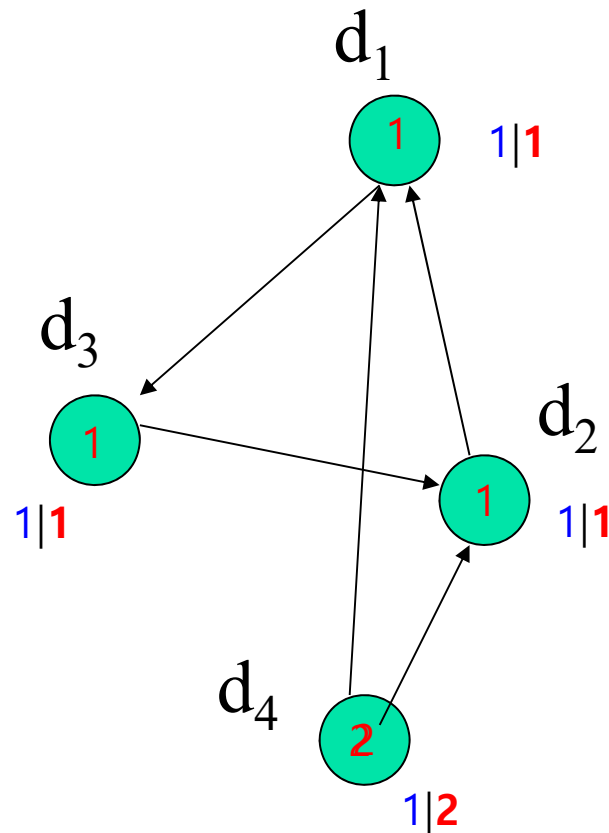


# The HITS Algorithm (Kleinberg 98)



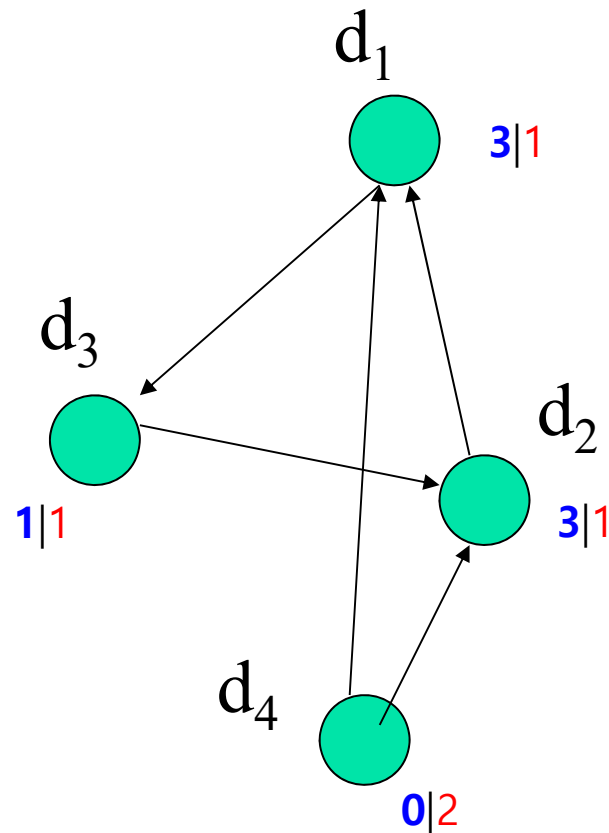
$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$
$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

# The HITS Algorithm (Kleinberg 98)



$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$
$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

# The HITS Algorithm (Kleinberg 98)



$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$

$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

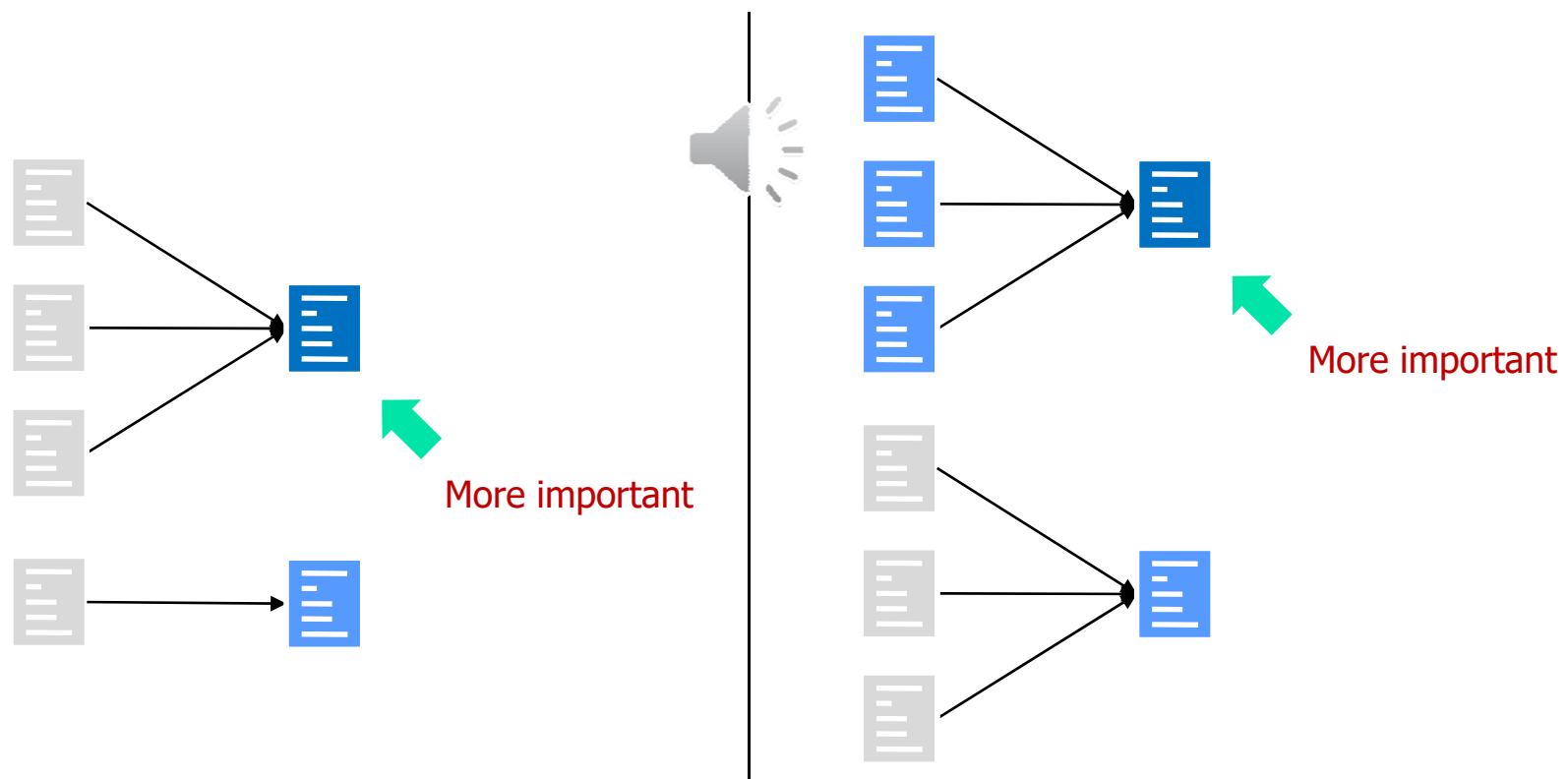
# PageRank: Capturing Page Popularity (Brin & Page'98)

---

- Intuitions
  - A page that is cited often can be expected to be more *important* (or authoritative) in general
- PageRank is essentially “citation counting”, but improves over simple counting
  - Consider “indirect citations” (being cited by a highly cited paper counts a lot...)
  - Smoothing of citations (every page is assumed to have a non-zero citation count)
- PageRank can also be interpreted as *random surfing* (thus capturing popularity)

# PageRank: Indirect Citations

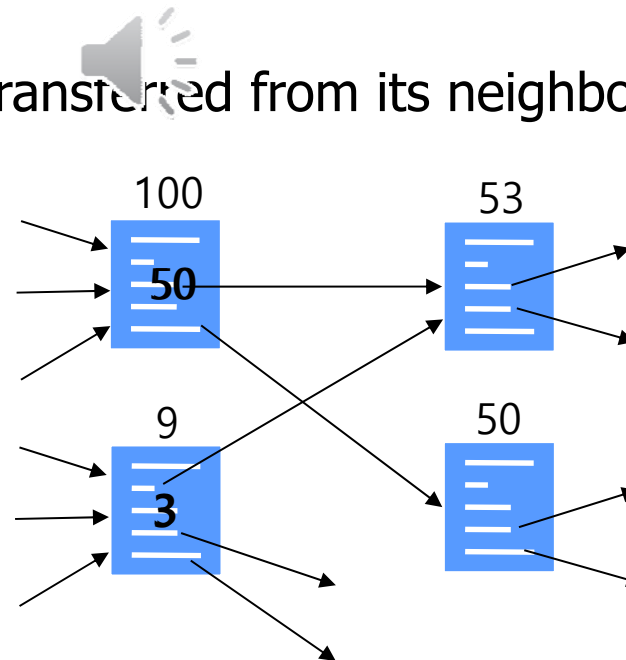
- A page that is *important* is...
  - Cited often by other pages
  - Cited often by other *important* pages





# PageRank: Simple Version

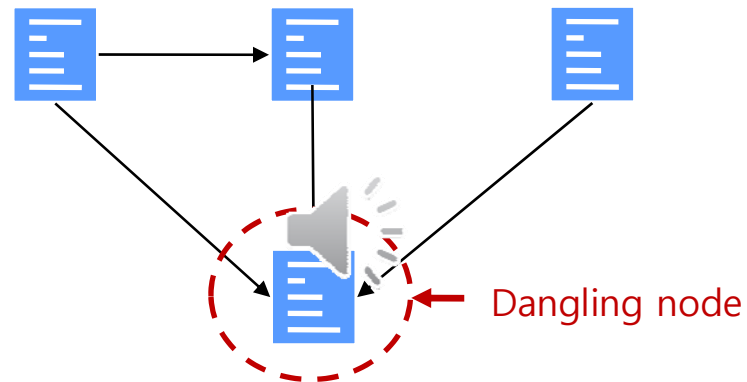
- Calculate importance score (authority score)
  - Initially, assign the same score to every page (e.g. 1)
  - For each page:
    - Transfer its score (divided equally) to its neighbors through out-links
    - Sum up the scores transferred from its neighbors through in-links
  - Iterate until...



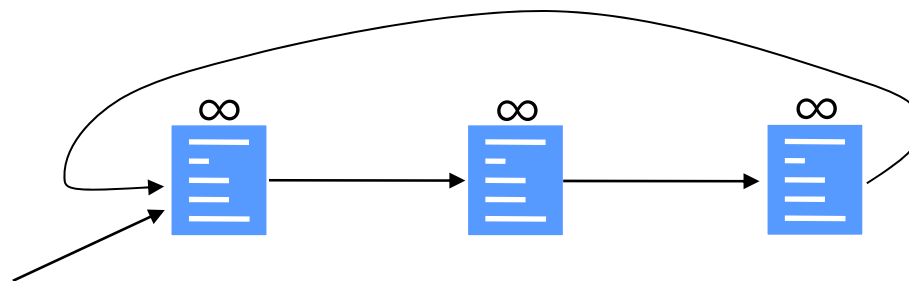
# PageRank: Simple Version

- Problems of the simple version

- Dangling nodes

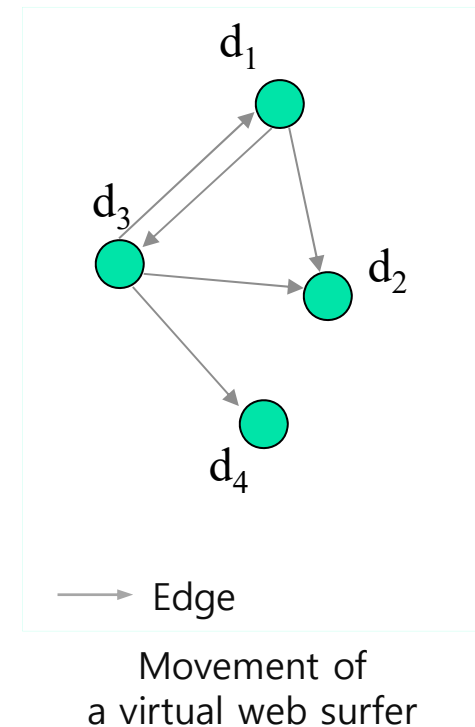


- Cyclic citation



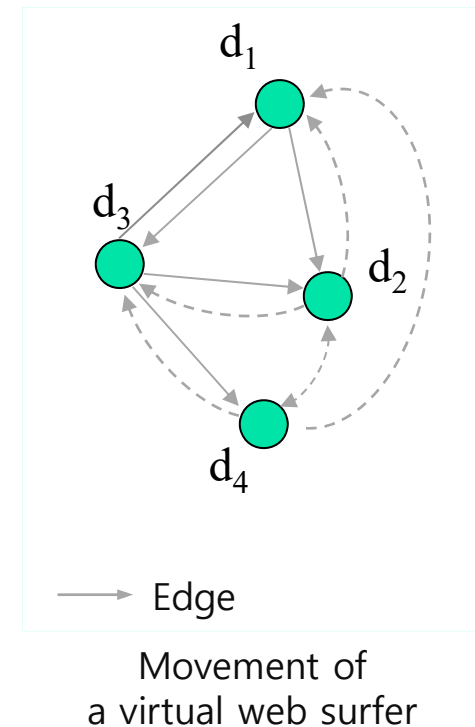
# PageRank: Random Surfer Model

- Random Surfer
  - Surfing the web by clicking on hyperlinks randomly
  - Or jump to a random page and *restart* surfing
- At any page,
  - With prob.  $\alpha$ , randomly jumping to a page (*restart*)
  - With prob.  $(1 - \alpha)$ , randomly picking a link to follow (*random walk*)



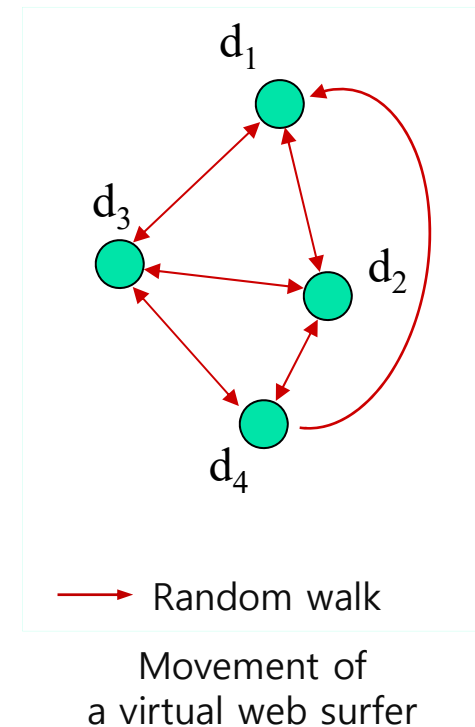
# PageRank: Random Surfer Model

- Random Surfer
  - Surfing the web by clicking on hyperlinks randomly
  - Or jump to a random page and *restart* surfing
- At any page,
  - With prob.  $\alpha$ , randomly jumping to a page (*restart*)
  - With prob.  $(1 - \alpha)$ , randomly picking a link to follow (*random walk*)



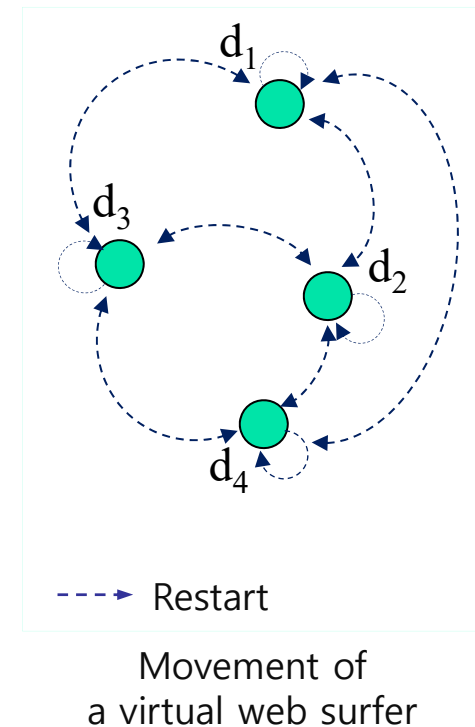
# PageRank: Random Surfer Model

- Random Surfer
  - Surfing the web by clicking on hyperlinks randomly
  - Or jump to a random page and *restart* surfing
- At any page,
  - With prob.  $\alpha$ , randomly jumping to a page (*restart*)
  - With prob.  $(1 - \alpha)$ , randomly picking a link to follow (*random walk*)



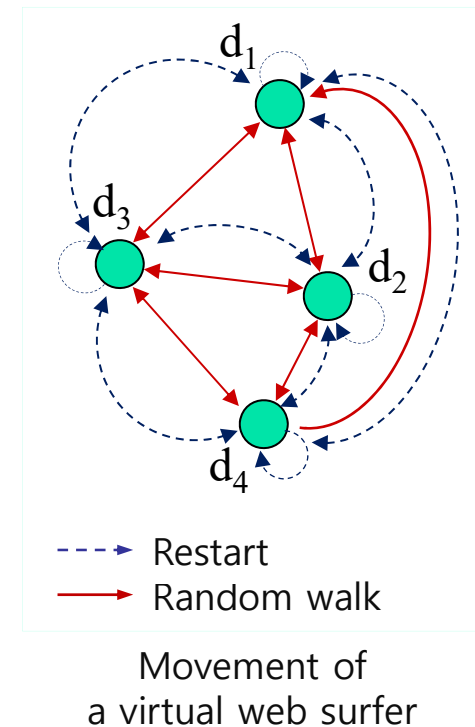
# PageRank: Random Surfer Model

- Random Surfer
  - Surfing the web by clicking on hyperlinks randomly
  - Or jump to a random page and *restart* surfing
- At any page,
  - With prob.  $\alpha$ , randomly jumping to a page (*restart*)
  - With prob.  $(1 - \alpha)$ , randomly picking a link to follow (*random walk*)

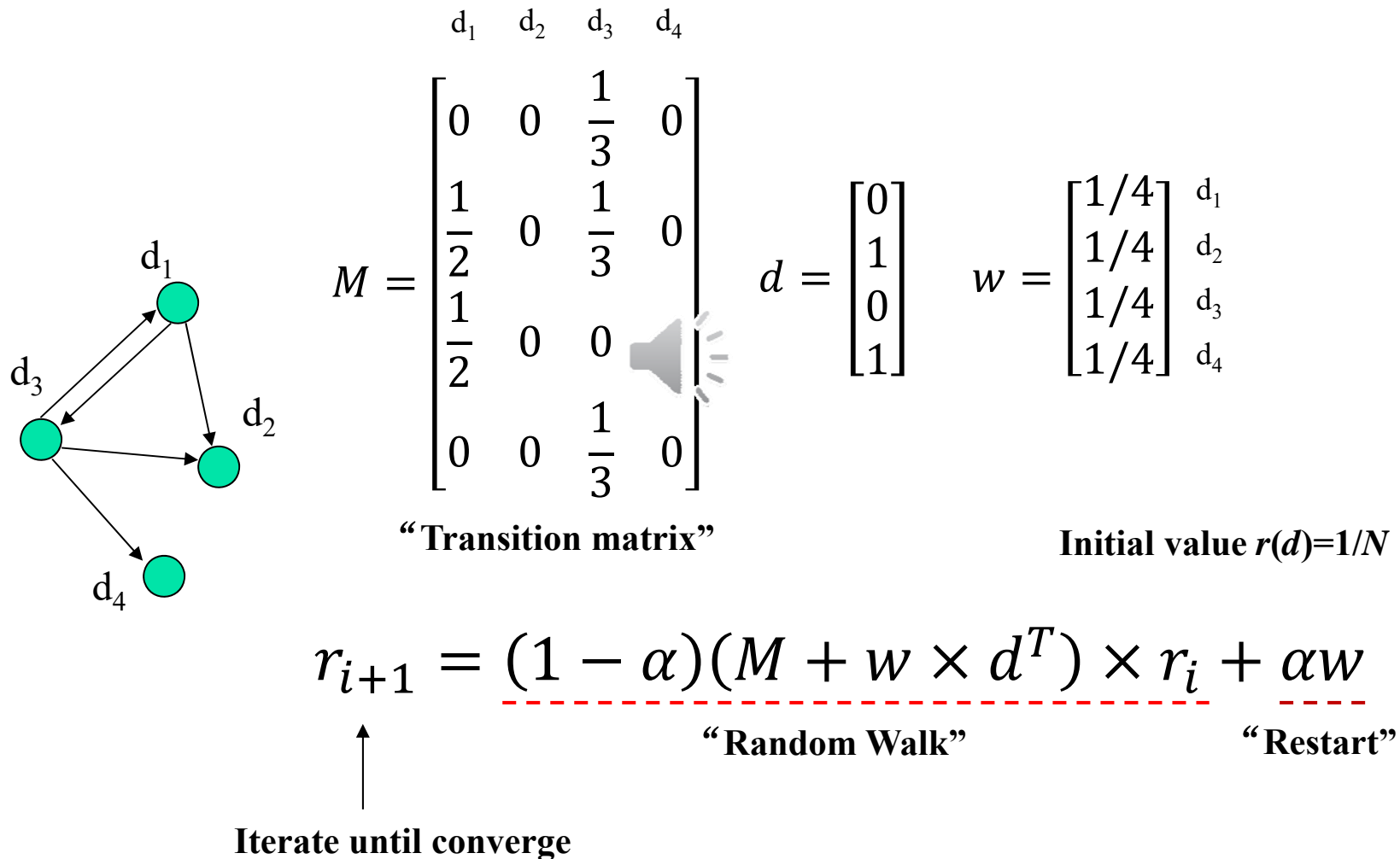


# PageRank: Random Surfer Model

- Random Surfer
  - Surfing the web by clicking on hyperlinks randomly
  - Or jump to a random page and *restart* surfing
- At any page,
  - With prob.  $\alpha$ , randomly jumping to a page (*restart*)
  - With prob.  $(1 - \alpha)$ , randomly picking a link to follow (*random walk*)

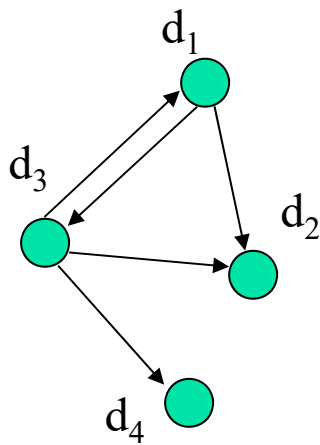


# The PageRank Algorithm (Brin & Page'98)





# The PageRank Algorithm (Brin & Page'98)



$$M + w \times d^T = \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{3} & \frac{1}{4} \end{bmatrix}$$

← No dangling node

Initial value  $r(d)=1/N$

$$r_{i+1} = (1 - \alpha) \underbrace{(M + w \times d^T)}_{\text{Solves dangling nodes problem}} \times r_i + \underbrace{\alpha w}_{\text{Solves cyclic citation problem}}$$

↑  
Iterate until converge

↑  
Solves dangling nodes problem

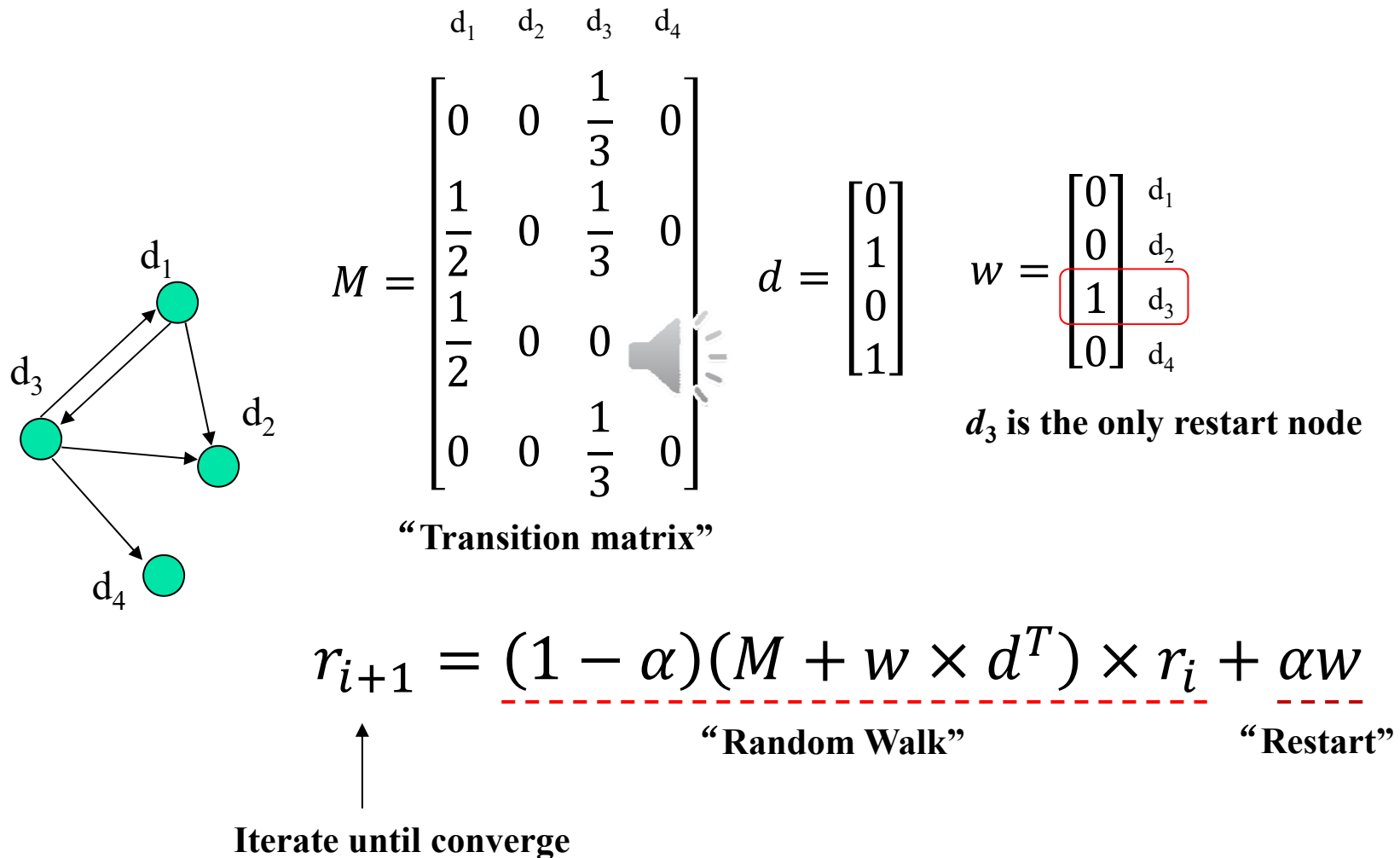
↑  
Solves cyclic citation problem

# RWR (Random Walk with Restart)

---

- Random walker on a graph
  - Start from the *restart nodes given*
  - Walk randomly to out-link nodes
  - Or jump randomly to restart nodes
- A random surfer model is a specialized case of RWR
  - Restart nodes include all the nodes

# RWR (Random Walk with Restart)




# Link-Based Object Classification (LBC)

---

- Predicting the category of an object based on its attributes, *its links, and the attributes of linked objects*
  - **Web**: Predict the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags, etc.
  - **Citation**: Predict the topic of a paper, based on word occurrence, citations, co-citations
  - **Epidemics**: Predict disease type based on characteristics of the patients infected by the disease

# Challenges in Link-Based Classification

---

- Labels of related objects tend to be correlated
- Collective classification: Explore such correlations and jointly infer the categorical values associated with the objects in the graph 
  - Ex: Classify related news items in Reuter data sets (Chak'98)
- Multi-relational classification is another solution for link-based classification

# Group Detection

---

- Cluster the nodes in the graph into groups that share common characteristics
  - **Web:** identifying communities
  - **Citation:** identifying research communities
- Methods
  - Hierarchical clustering
  - Blockmodeling of SNA
  - Spectral graph partitioning
  - Stochastic blockmodeling
  - Multi-relational clustering

# Entity Resolution

---

- Predicting when two objects are the same, based on their attributes and their links
- Also known as: deduplication, reference reconciliation, co-reference resolution, object consolidation
- Applications
  - **Web:** predict when two sites are mirrors of each other
  - **Citation:** predicting when two citations are referring to the same paper
  - **Epidemics:** predicting when two disease strains are the same
  - **Biology:** learning when two names refer to the same protein



# Entity Resolution Methods

---

- Earlier viewed as pair-wise resolution problem: resolved based on the similarity of their attributes
- Importance at considering links
  - Coauthor links in bib data, hierarchical links between spatial references, co-occurrence links between name references in documents
- Use of links in resolution
  - Collective entity resolution: one resolution decision affects another if they are linked
    - Propagating evidence over links
  - Probabilistic models interact with different entity recognition decisions




# Link Prediction

---

- Predict whether a link exists between two entities, based on attributes and other observed links
- Applications
  - **Web**: predict if there will be a link between two pages
  - **Citation**: predicting if a paper will cite another paper
  - **Epidemics**: predicting who a patient's contacts are
- Methods
  - Often viewed as a binary classification problem
  - Local conditional probability model, based on structural and attribute features
  - Difficulty: sparseness of existing links
  - Collective prediction, e.g., Markov random field model

# Link Cardinality Estimation

---

- Predicting the number of links to an object
  - **Web**: predict the authority of a page based on the number of in-links; identifying hubs based on the number of out-links
  - **Citation**: predicting the impact of a paper based on the number of citations 
  - **Epidemics**: predicting the number of people that will be infected based on the infectiousness of a disease
- Predicting the number of objects reached along a path from a *specific object*
  - **Web**: predicting number of pages retrieved by crawling a site
  - **Citation**: predicting the number of citations of a particular author in a specific journal

# Subgraph Discovery

---

- Find characteristic subgraphs
  - Focus of graph-based data mining
- Applications
  - **Biology:** protein structure discovery
  - **Communications:** legitimate vs. illegitimate groups
  - **Chemistry:** chemical substructure discovery
- Methods
  - Subgraph pattern mining
- Graph classification
  - Classification *based on subgraph pattern analysis*

# Metadata Mining

---

- Schema mapping, schema discovery, schema reformulation
  - **cite** - matching between two bibliographic sources
  - **web** - discovering schema from unstructured or semi-structured data
  - **bio** - mapping between two medical ontologies

# Thanks!

- Jiwon Hong ([nowiz@hanyang.ac.kr](mailto:nowiz@hanyang.ac.kr))



# Ref: Mining on Social Networks

---

- D. Liben-Nowell and J. Kleinberg. The Link Prediction Problem for Social Networks. CIKM'03
- P. Domingos and M. Richardson, Mining the Network Value of Customers. KDD'01
- M. Richardson and P. Domingos, Mining Knowledge-Sharing Sites for Viral Marketing. KDD'02
- D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the Spread of Influence through a Social Network. KDD'03.
- P. Domingos, Mining Social Networks for Viral Marketing. IEEE Intelligent Systems, 20(1), 80-82, 2005.
- S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine. WWW7.
- S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, Mining the link structure of the World Wide Web. IEEE Computer'99
- D. Cai, X. He, J. Wen, and W. Ma, Block-level Link Analysis. SIGIR'2004.