Information Networks and the WWW

Link Analysis

Overview

- Intro to Web Search
- Crawling
- Web Link Analysis
- Beyond Page Rank
  - Trust Rank
  - HillTop Algorithm
  - Sim Rank
- Link Analysis in other scenarios

Page Rank - Advantages

- Vectorized system of equations fast to compute
- Guaranteed to converge to a unique solution
- Ranks can be pre-computed during indexing and reused during query time
- Ranks are robust and stable as in edges to a page are harder to manipulate than out edges
- Conforms with the intuitive notion of importance of entities from the real world

Page Rank - Disadvantages

- Prone to spamming, as it considers only the connections of node rather than its content
- A page can get high rank by connecting a lot of trivial (possibly dummy) pages
- The basic PageRank system assumes a static system; no modification in adjacency matrix allowed during computation
- For dynamic systems, any modification requires all ranks to be re computed

Page Rank in Web Search

- Page Rank started by Google and used in many other Search engines
- Currently Google may be using a different technique – most probably combination of various algorithms
- An effective algorithm needs to combine multiple methods
  - Anchor Text (We can give more weights to those links)
  - User feedback
- Must guard against Web sites that may try to get on top of Search results
  - Pages created to fool search engines – SPAMDEXING.
  - Search engine business depends on successful filtering of SPAM pages

Overview

- Intro to Web Search
- Crawling
- Web Link Analysis
- Beyond Page Rank
  - Trust Rank
  - HillTop Algorithm
  - SimRank
- Link Analysis in other scenarios

Some Spamming Techniques

- Changing a color scheme for keyword stuffing: e.g. white text on white background
- Link farms: Creating a number of bogus web pages that link to one page in order to give it a better rank
- Honey Pot: Provide some valuable content but contain links to SPAM pages
- Scraper sites: Scrape content from search engines and other websites
- Article spinning: Rewriting existing articles to escape duplicate content penalty
- Cloaking: Serving the page differently to the crawler than to the humans

How do we filter out SPAM?

- How do we separate the "good" ones from the "bad" ones?
- It turns out its hard to do it automatically
- The most reliable way is to use human experts but how do we do that for billions of pages?
- Is there a way to conclude something based on a small set of pages reviewed by experts?
- An approach – Trust Rank [3]

Problem Definition

- How can we semi-automatically estimate which pages are good and which ones are bad (SPAM) provided that we have a limited number of experts?
- Can we reliably say which ones are probably good and/or probably bad based on a small "seed" of pages reviewed by experts?
- How can we do it effectively and efficiently?

Oracle

- The notion of human checking of a web page is represented by Oracle function:

$$O(p) = \begin{cases} 0 & \text{if } p \text{ is bad,} \\ 1 & \text{if } p \text{ is good.} \end{cases}$$

- Oracle invocations are expensive, one should strive to minimize them
- Important empirical observation for trust: *approximate isolation* of the good set
  - Good pages rarely link to bad ones
  - The converse does not hold

Trust Function

- To evaluate the pages without calling O, it is necessary to estimate the probability that p is good
- The Trust function yields a range of values between 0 (bad) and 1 (good)
- Ideally,

$$T(p) = \Pr[O(p) = 1].$$

- This is hardly ever true in practice
- A relaxed constraint is orderedness by pair, so that we can display search results based on that order

$$T(p) < T(q) \Leftrightarrow \Pr[O(p) = 1] < \Pr[O(q) = 1].$$
$$T(p) = T(q) \Leftrightarrow \Pr[O(p) = 1] = \Pr[O(q) = 1].$$

Trust Function

- Another method of relaxing the requirements of T is introducing a **threshold** value

$$T(p) > \delta \Leftrightarrow O(p) = 1$$

- If a page receives a score above δ we know it is good.
- Otherwise, we cannot say anything
- This does not necessarily provide ordering based on the likelihood of being good

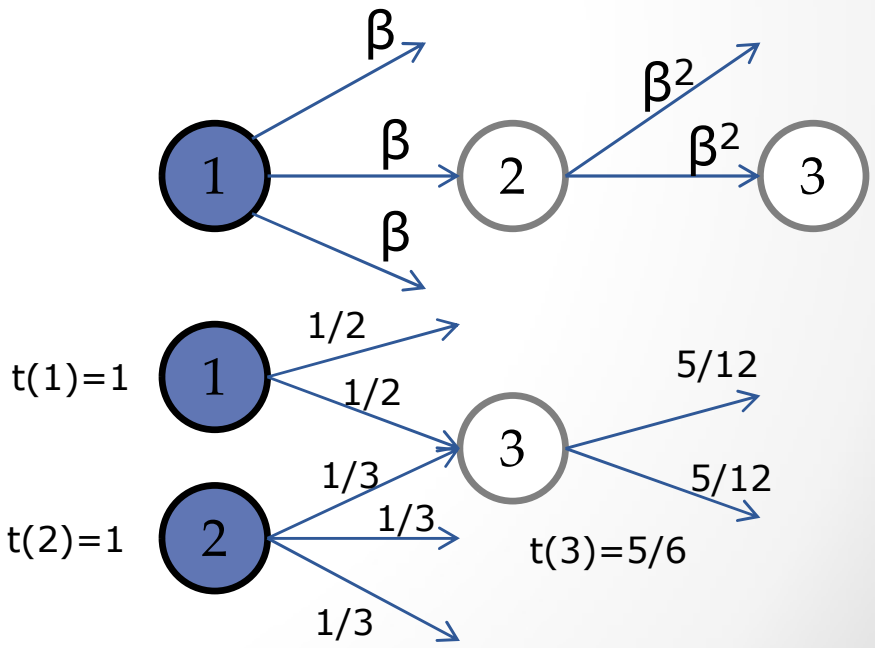
Seed Selection

- Random selection is simplest but it may hinder TrustRank effectiveness
- Oracle invocations are expensive, because they require human effort
- Chosen pages should be useful in identifying additional good pages
- Seed set should be reasonably small to limit oracle invocations
- Trust flows out of the good seed pages, so give preference to pages which reach many other pages
- Select pages based on the number of outlinks
- Scheme closely related to PageRank
- Difference is that *trust* depends on outlinks not inlinks

Trust Attenuation

- We cannot be absolutely sure that pages reachable from good seeds are indeed good
- Further away we are from good seed, less certain we are that a page is good

- Trust dampening
  β – dampening factor
- Trust splitting
- Trust combination



Further Improvements

- During the seed selection it is not necessary to order all pages using inversed PageRank
- If we already have PageRank score, we can choose subset of highly ranked pages
  - Users will be more interested in those pages anyway
- Anti-Trust Rank: The closer a site is to spam resources, the more likely it is to be spam as well.

Overview

- Intro to Web Search
- Crawling
- Web Link Analysis
- Beyond Page Rank
  - Trust Rank
  - HillTop Algorithm
  - SimRank
- Link Analysis in other scenarios

Hilltop Algorithm

- The Hilltop algorithm is an algorithm used to find documents relevant to a particular keyword topic in news search [4]
- Incorporated in Google
- Assumption: The number and quality of the sources referring to a page are a good measure of the page's quality.
- Novelty: Only considering "expert" sources - pages that have been created with the specific purpose of directing people towards resources.

Hilltop Algorithm

- In response to a query:
  i. Compute a list of the most relevant experts on the query topic.
  ii. Identify relevant links within the selected set of experts and follow them to identify target web pages.
  iii. The targets are then ranked according to the number and relevance of non-affiliated experts that point to them.
  iv. Thus, the score of a target page reflects the collective opinion of the best independent experts on the query topic.
  v. When such a pool of experts is not available, Hilltop provides no results.
- Thus, Hilltop is tuned for result accuracy and not query coverage.
- Algorithm consists of two broad phases:
  i. Expert Lookup
  ii. Target Ranking

Expert Lookup

- An expert page is a page that is about a certain topic and has links to many non-affiliated pages on that topic.
  - Two pages are non-affiliated conceptually if they are authored by authors from non-affiliated organizations.
- In a pre-processing step, a subset of the pages crawled by a search engine are identified as experts
- Given an input query, a lookup is done on the expert-index to find and rank matching expert pages.
- This phase computes the best expert pages on the query topic as well as associated match information.

Target Ranking

- A page is an authority on the query topic if and only if some of the best experts on the query topic point to it.
- Of course in practice some expert pages may be experts on a broader or related topic.
  - If so, only a subset of the hyperlinks on the expert page may be relevant.
- By combining relevant out-links from many experts on the query topic we can find the pages that are most highly regarded by the community of pages related to the query topic
- Given the top ranked matching expert-pages and associated match information, we select a subset of the hyperlinks within the expert-pages.
- Specifically, we select links that we know to have all the query terms associated with them.
  - This implies that the link matches the query.
- With further connectivity analysis on the selected links we identify a subset of their targets as the top-ranked pages on the query topic.
- The targets identified are those that are linked to by *at least two* non-affiliated expert pages on the topic.
- The targets are ranked by a ranking score which is computed by combining the scores of the experts pointing to the target.

Overview

- Intro to Web Search
- Crawling
- Web Link Analysis
- Beyond Page Rank
  - Trust Rank
  - HillTop Algorithm
  - SimRank
- Link Analysis in other scenarios

Measuring Similarity

- The problem of measuring similarity of objects (nodes in a graph) arises in many applications
- The approach should be applicable in any domain with object to object relationships
- Metadata used to measure similarity between objects are often hard to determine and quantify in practice
- Contextual information may be used for the purpose
  - **Two objects are similar if they are related to similar objects**
  - Easier to determine in practice
- SimRank [5] follows the above paradigm to measure similarity between entities
- SimRank is efficient as well
  - For a network of size N, we require N^2 similarity score, one per each pair of objects

SimRank

- We model objects and relationships as a directed graph $G = (V, E)$
- For a node *v* in a graph, we denote by $I(v)$ and $O(v)$ the set of in-neighbors and out-neighbors
- Basic SimRank Equation
  - If $a = b$ then $s(a, b)$ is defined to be 1. Otherwise,

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)||I(b)|} \sum_{j=1} s(I_i(a), I_j(b))$$

    where *C* is a constant between 0 and 1
  - Set $s(a, b) = 0$ when $I(a) = \emptyset$ or $I(b) = \emptyset$

SimRank: How to compute?

- Computing SimRank — Naive Method
  - $R_0(a,b)$ is initialized with the lower bound on the $s(a,b)$

$$R_0(a,b) = \begin{cases} 0 & (\text{if } a \neq b) \\ 1 & (\text{if } a = b) \end{cases}$$

  - **To compute $R_{k+1}(a,b)$ from $R_k(\cdot,\cdot)$:**

$$R_{k+1}(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k(I_i(a), I_j(b)) \qquad \text{when } a \neq b$$

$$R_{k+1}(a,b) = 1 \qquad a = b$$

SimRank in Bi-partite Network

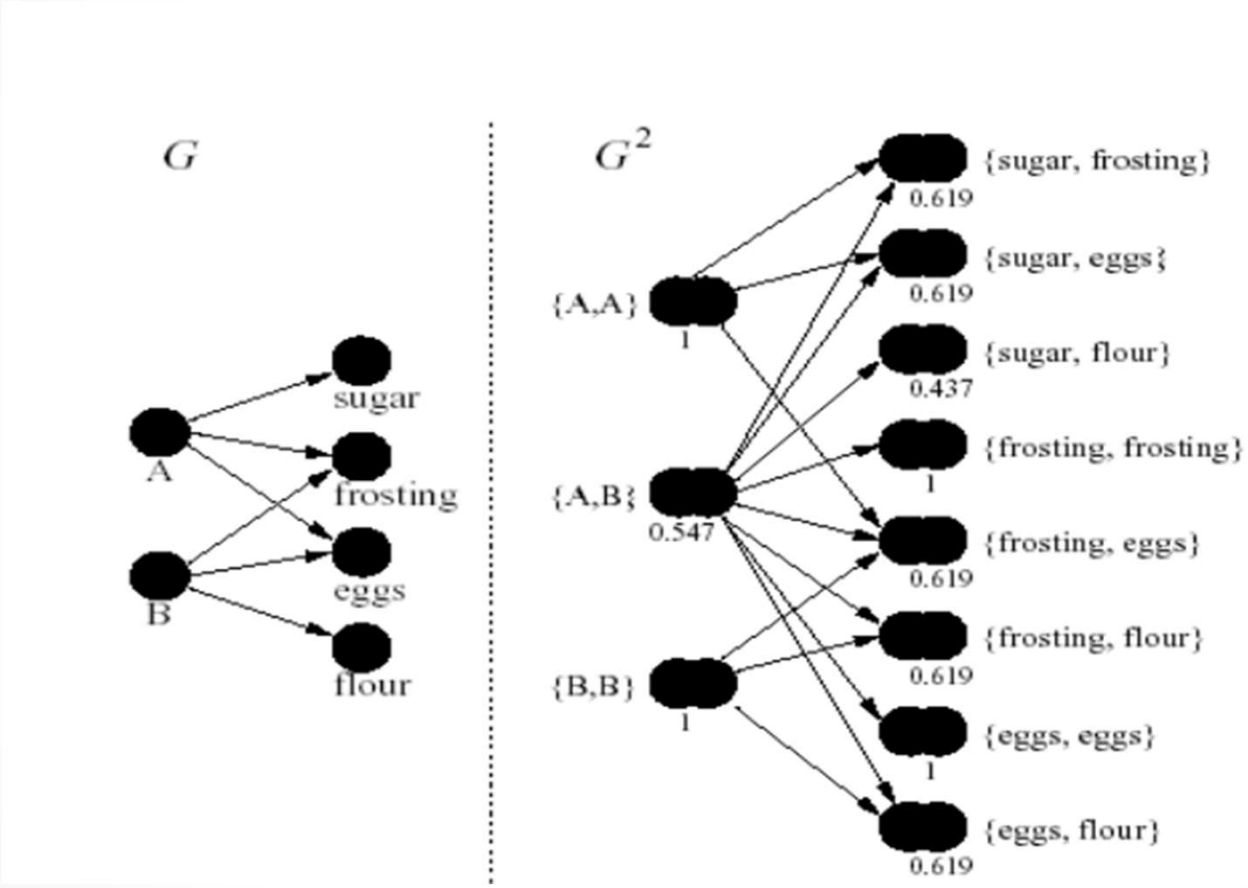
- In a heterogeneous network of users and products, the similarity of products and users are mutually-reinforced
  - two users can be considered similar if they buy similar products
  - two products can be considered similar if they are bought by similar users
- Similarity between two distinct users can be expressed as:

$$s(u_1, u_2) = \frac{C_1}{|O(u_1)|.|O(u_2)|} \sum_{i=1}^{|O(u_1)|} \sum_{j=1}^{|O(u_2)|} s(O_i(u_1), O_j(u_2))$$

- Similarity between two distinct products can be expressed as:

$$s(p_1, p_2) = \frac{C_2}{|I(p_1)|.|I(p_2)|} \sum_{i=1}^{|I(p_1)|} \sum_{j=1}^{|I(p_2)|} s(I_i(p_1), I_j(p_2))$$

SimRank in Bi-partite Network: Example



Overview

- Intro to Web Search
- Crawling
- Web Link Analysis
- Beyond Page Rank
- Link Analysis in other scenarios

Citation Analysis

- Impact factor for a scientific journal is defined to be the average number of citations received by a paper in the given journal over the past two years.

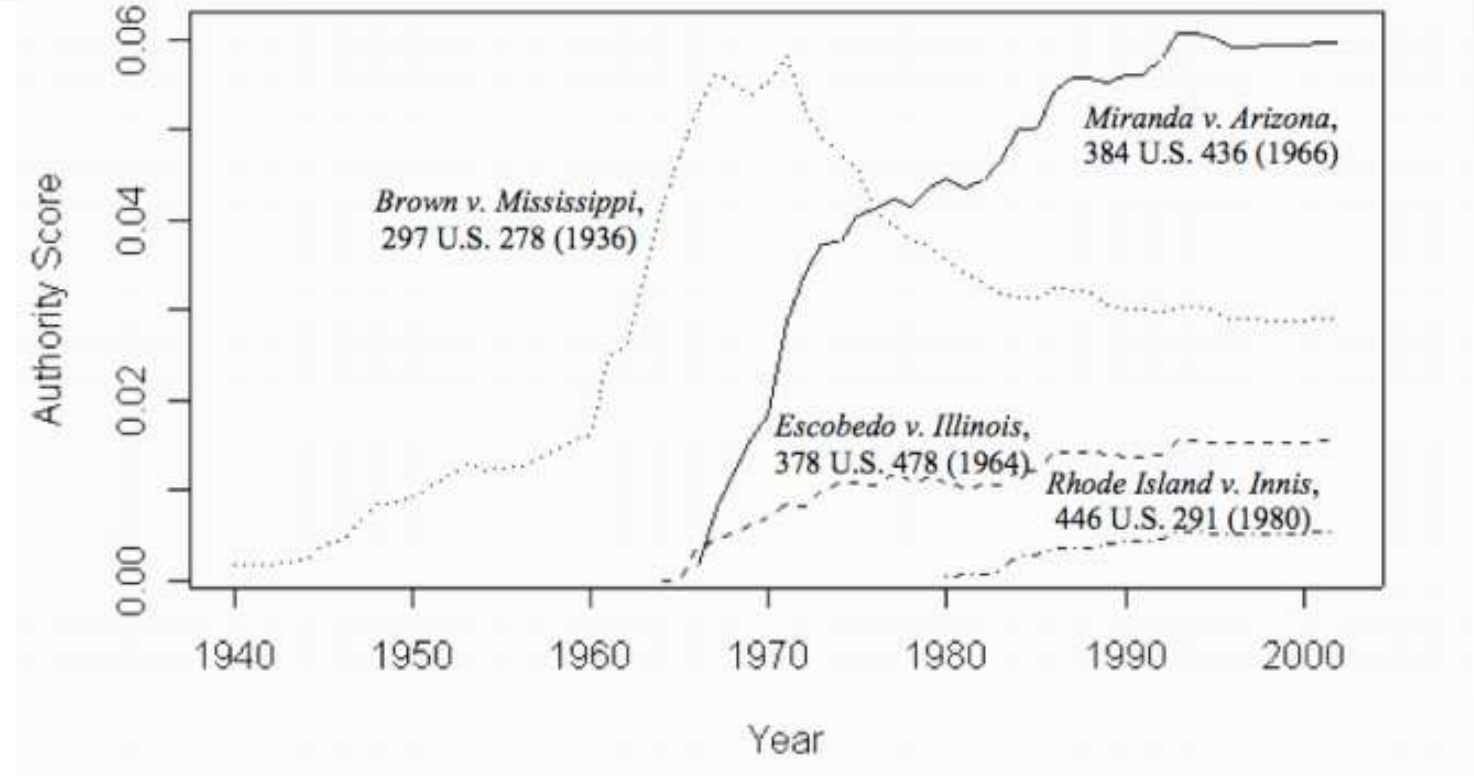
- This type of voting by in-links can thus serve as a proxy for the collective attention that the scientific community pays to papers published in the journal.

- In the 1970s, Pinski and Narin [6] extended the impact factor by taking into account the idea that not all citations should be counted equally — rather, citations from journals that are themselves high-impact should be viewed as more important

Citation Analysis in US Supreme Court

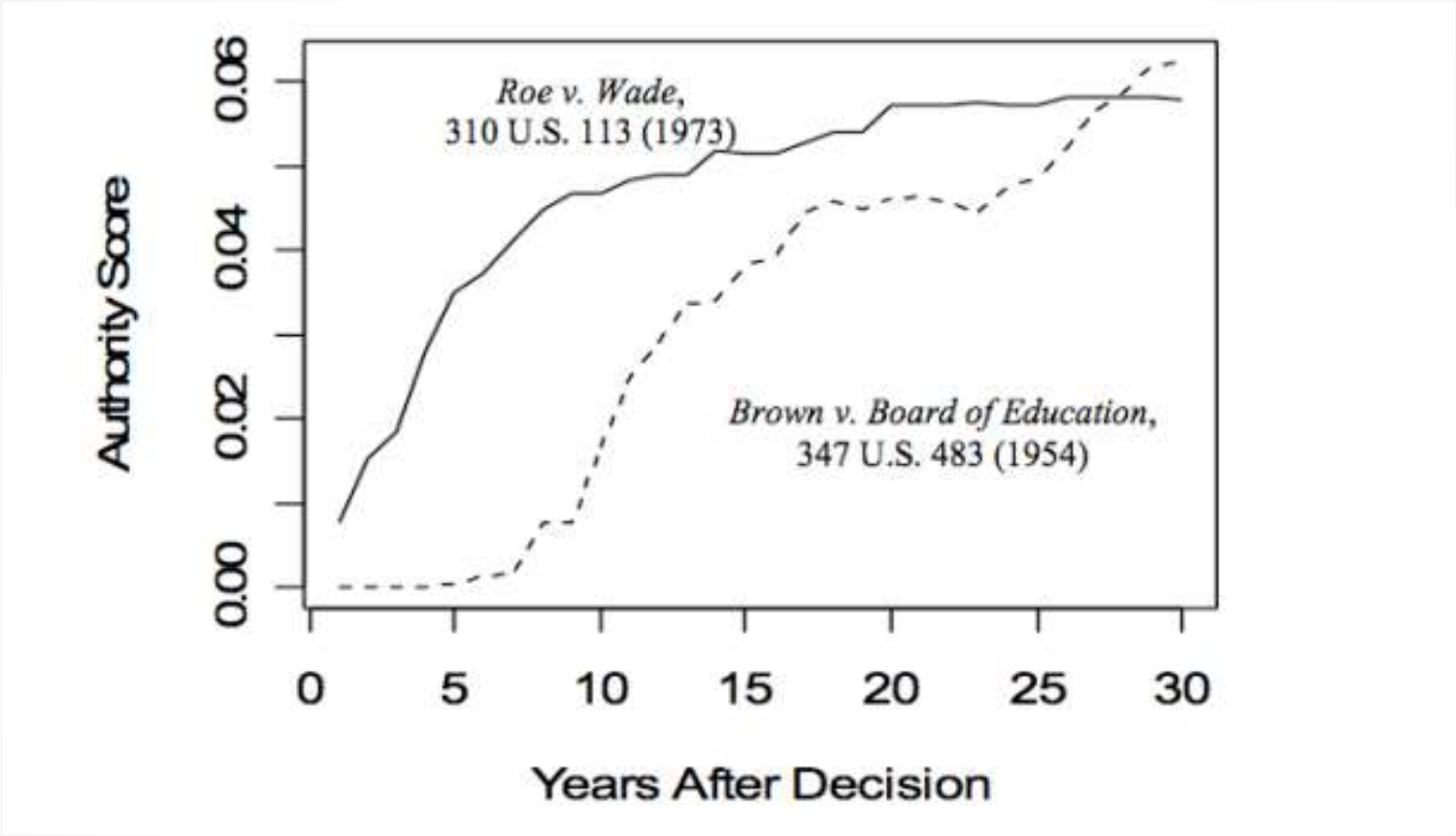
- Citations are crucial in legal writing, to ground a decision in precedent and to explain the relation of a new decision to what has come before.
- Link analysis in this context can help in identifying cases that play especially important roles in the overall citation structure.
- In one example of this style of research, Fowler and Jeon [7] applied hub and authority measures to the set of all U.S. Supreme Court decisions, a collection of documents that spans more than two centuries.
- They found that the set of Supreme Court decisions with high authority scores in the citation network align well with the more qualitative judgments of legal experts about the Court's most important decisions.

Citation Analysis in US Supreme Court



Authority for a particular topic can change over long time periods.

Citation Analysis in US Supreme Court



Significant decisions can vary widely in the rate at which they acquire authority.

References

1. Jon Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999
2. Brin & Page. The anatomy of a large-scale hypertextual Web search engine The anatomy of a large-scale hypertextual Web search engine In Proceedings of 7th International World Wide Web Conference, 1998.
3. Gyongyi, Zoltan; Garcia-Molina, Hector. Combating Web Spam with TrustRank. Proceedings of the 30th VLDB Conference, 2004.
4. Krishna Bharat and George A. Mihaila. When experts agree: Using non-affiliated experts to rank popular topics. In Proceedings 10th International World Wide Web Conference, 2001
5. Glen Jeh, Jennifer Widom. SimRank: A Measure of Structural-Context Similarity. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002
6. Gabriel Pinski and Francis Narin. Citation influence for journal aggregates of scientic publications: Theory, with application to the literature of physics. Information Processing and Management, 12:297-312, 1976
7. James H. Fowler and Sangick Jeon. The authority of Supreme Court precedent. Social Networks, 30:16-30, 2008

Reading

1. David Easley and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. https://www.cs.cornell.edu/home/kleinber/networks-book/
   - Chapter 14
2. Social Network Analysis. Tanmoy Chakraborty.
   - Chapter 4