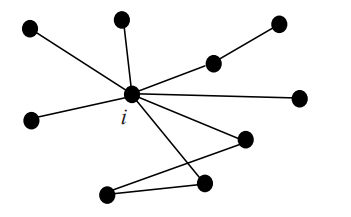
**Social Network Analysis**

Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships. The interactions and relationships can be represented with a network or graph, where each vertex (or node) represents an actor and each link represents a relationship. From the network we can study the properties of its structure, and the role, position and prestige of each social actor. We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors. Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. Many of the results from social networks can be adapted and extended for use in the Web context. The ideas from social network analysis are indeed instrumental to the success of Web search engines. In this section, we introduce two types of social network analysis, centrality and prestige, which are closely related to hyperlink analysis and search on the Web. Both centrality and prestige are measures of degree of prominence of an actor in a social network.

**Centrality**

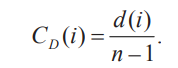
Centrality Important or prominent actors are those that are linked or involved with other actors extensively. In the context of an organization, a person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts. The links can also be called ties. A central actor is one involved in many ties. Figure shows a simple example using an undirected graph. Each node in the social network is an actor and each link indicates that the actors on the two ends of the link communicate with each other. Intuitively, we see that the actor i is the most central actor because he/she can communicate with most other actors.



* **Degree Centrality**

Degree Centrality Central actors are the most active actors that have most links or ties with other actors. Let the total number of actors in the network be n.

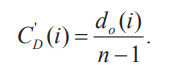
**Undirected Graph:** In an undirected graph, the degree centrality of an actor i (denoted by CD(i)) is simply the node degree (the number of edges) of the actor node, denoted by d(i), normalized with the maximum degree, n−1.



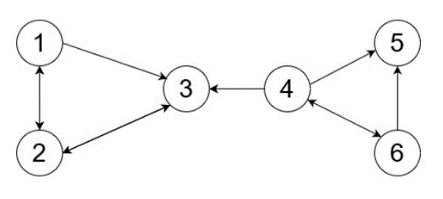
The value of this measure ranges between 0 and 1 as n−1 is the maximum value of d(i).

C0(1)=2/5=0.4

**Directed Graph**: In this case, we need to distinguish in-links of actor i (links pointing to i), and out-links (links pointing out from i). The degree centrality is defined based on only the out-degree (the number of out-links or edges), do(i).



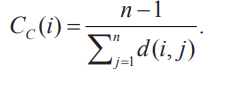
C’0(1)=2/5=0.4



* **Closeness Centrality**

This view of centrality is based on the closeness or distance. The basic idea is that an actor xi is central if it can easily interact with all other actors. That is, its distance to all other actors is short. Thus, we can use the shortest distance to compute this measure. Let the shortest distance from actor i to actor j be d(i, j) (measured as the number of links in a shortest path).

**Undirected Graph**: The closeness centrality CC(i) of actor i is defined as



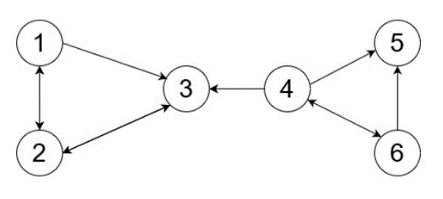
The value of this measure also ranges between 0 and 1 as n−1 is the minimum value of the denominator, which is the sum of the shortest distances from i to all other actors. Note that this equation is only meaningful for a connected graph.

Cc(1)=5/1+1+2+3+3=5/10=0.5

**Directed Graph:** The same equation can be used for a directed graph. The distance computation needs to consider directions of links or edges.

C’c(1)=5/1+1+100+100+100

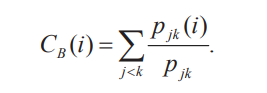
C’c(3)=5/2+1+100+100+100



* **Betweenness Centrality**

If two non-adjacent actors j and k want to interact and actor i is on the path between j and k, then i may have some control over their interactions. Betweenness measures this control of i over other pairs of actors. Thus, if i is on the paths of many such interactions, then i is an important actor.

**Undirected Graph:** Let pjk be the number of shortest paths between actors j and k. The betweenness of an actor i is defined as the number of shortest paths that pass i (denoted by pjk(i), j ≠ i and k ≠ i) normalized by the total number of shortest paths of all pairs of actors not including i:



**Directed Graph:** The same equation can be used but must be multiplied by 2 because there are now (n−1)(n−2) pairs considering a path from j to k is different from a path from k to j. Likewise, pjk must consider paths from both directions.

**Betweeness=3**

**1—2--------0**

**1—4--------1**

**1—5---------1**

**1—6---------1**

**2—4----------1**

**2—5--------1**

**2—6---------1**

**4—5---------0**

**4—6---------0**

**5—6---------0**

**Value is 6**

**Prestige**

Prestige is a more refined measure of prominence of an actor than centrality as we will see below. We need to distinguish between ties sent (out-links) and ties received (in-links). A prestigious actor is defined as one who is object of extensive ties as a recipient. In other words, to compute the prestige of an actor, we only look at the ties (links) directed or pointed to the actor (in-links). Hence, the prestige cannot be computed unless the relation is directional or the graph is directed. The main difference between the concepts of centrality and prestige is that centrality focuses on out-links while prestige focuses on in-links. We define three prestige measures. The third prestige measure (i.e., rank prestige) forms the basis of most Web page link analysis algorithms, including PageRank and HITS.

**Degree Prestige**

Degree Prestige Based on the definition of the prestige, it is clear that an actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor i (denoted by PD(i)) is its in-degree,

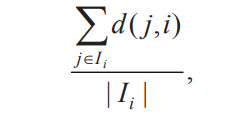


where dI(i) is the in-degree of i (the number of in-links of i) and n is the total number of actors in the network. As in the degree centrality, dividing by n – 1 standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor i.

Pd(1)=1/5=0.2

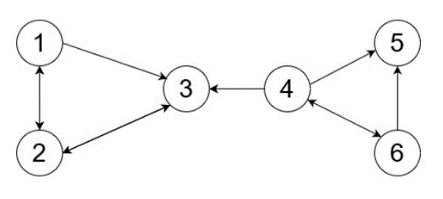
**Proximity Prestige**

The degree index of prestige of an actor i only considers the actors that are adjacent to i. The proximity prestige generalizes it by considering both the actors directly and indirectly linked to actor i. That is, we consider every actor j that can reach i, i.e., there is a directed path from j to i. Let Ii be the set of actors that can reach actor i, which is also called the influence domain of actor i. The proximity is defined as closeness or distance of other actors to i. Let d(j, i) denote the shortest path distance from actor j to actor i. Each link has the unit distance. To compute the proximity prestige, we use the average distance, which is



I1={2,3,4,6}

Pp(1)=1+2+3+4/4=2.5

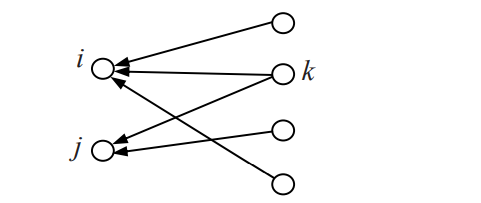


**Co-Citation and Bibliographic**

Co-Citation and Bibliographic Coupling Another area of research concerned with links is the citation analysis of scholarly publications. A scholarly publication usually cites related prior work to acknowledge the origins of some ideas in the publication and to compare the new proposal with existing work. Citation analysis is an area of bibliometric research, which studies citations to establish the relationships between authors and their work. When a publication (also called a paper) cites another publication, a relationship is established between the publications. Citation analysis uses these relationships (links) to perform various types of analysis. A citation can represent many types of links, such as links between authors, publications, journals and conferences, and fields, or even between countries. We will discuss two specific types of citation analysis, co-citation and bibliographic coupling.

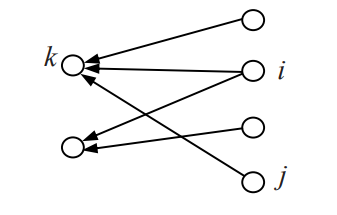
**Co-Citation**

Co-citation is used to measure the similarity of two papers (or publications). If papers i and j are both cited by paper k, then they may be said to be related in some sense to each other, even though they do not directly cite each other. Figure shows that papers i and j are co-cited by paper k. If papers i and j are cited together by many papers, it means that i and j have a strong relationship or similarity. The more papers they are cited by, the stronger their relationship is.

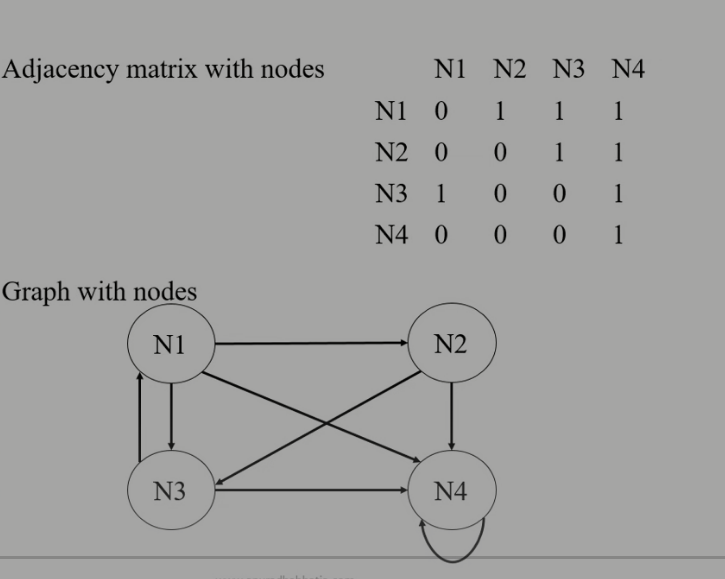


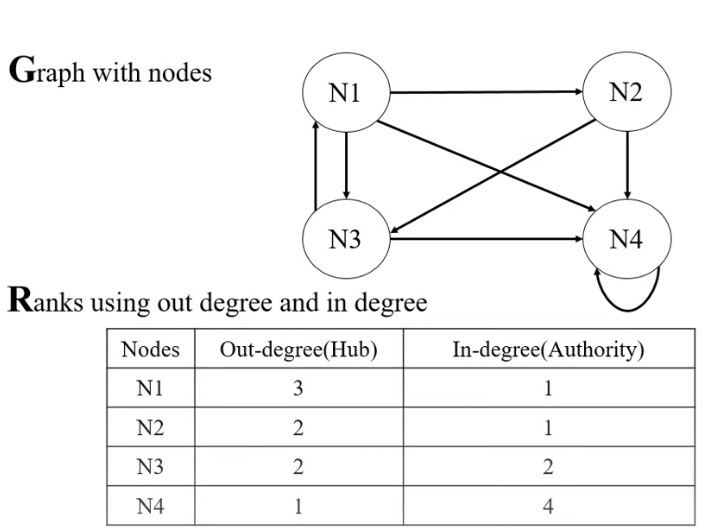
**Bibliographic Coupling**

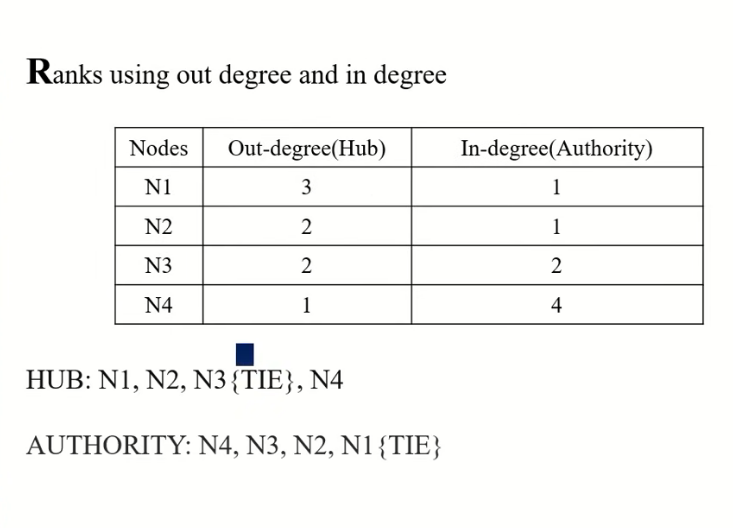
Bibliographic Coupling Bibliographic coupling operates on a similar principle, but in a way it is the mirror image of co-citation. Bibliographic coupling links papers that cite the same articles so that if papers i and j both cite paper k, they may be said to be related, even though they do not directly cite each other. The more papers they both cite, the stronger their similarity is. Fig. 7.4 shows both papers i and j citing (referencing) paper k.

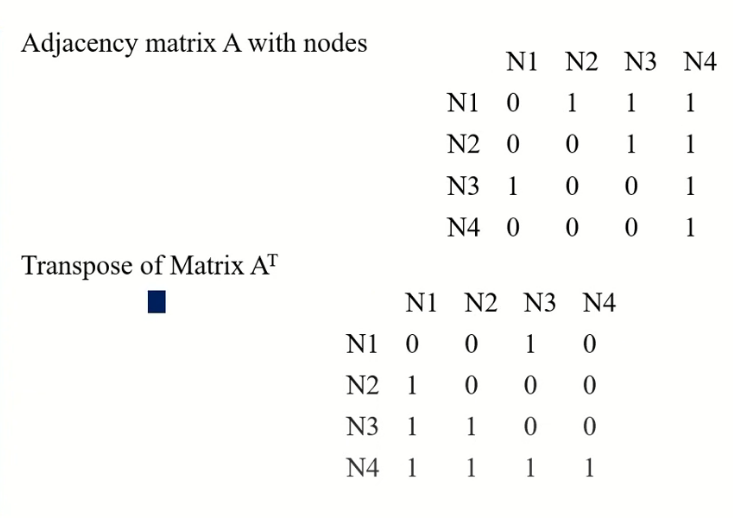


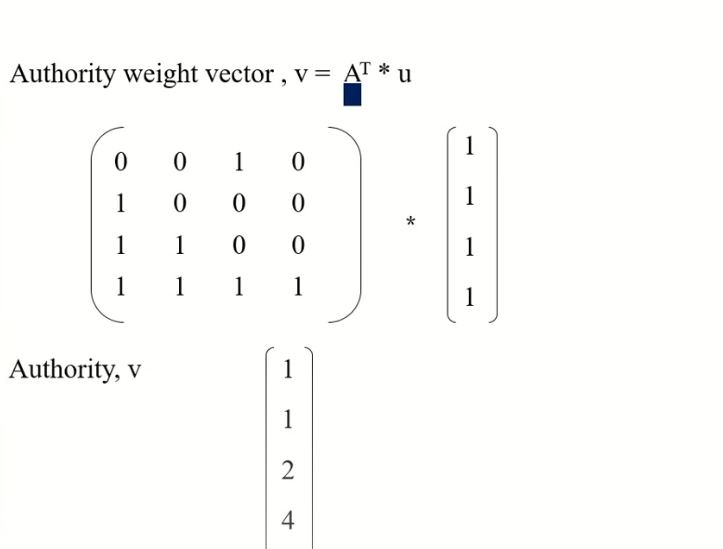
**Hyperlink-Induced Topic Search (HITS)** (also known as Hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

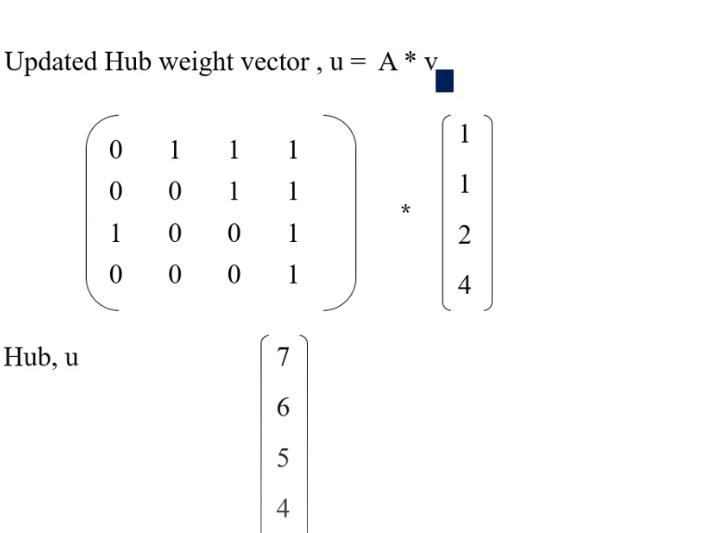


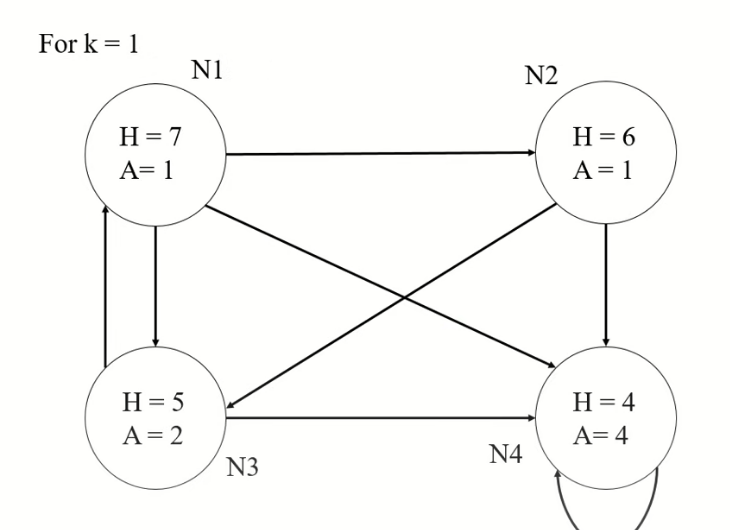


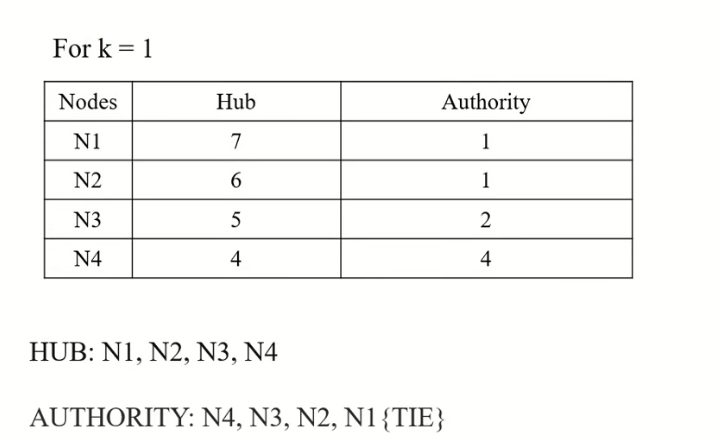


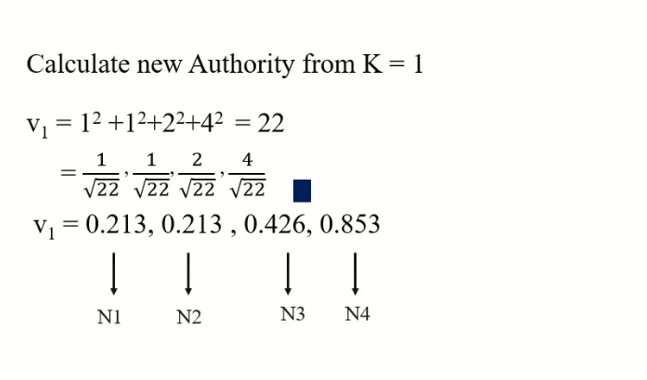


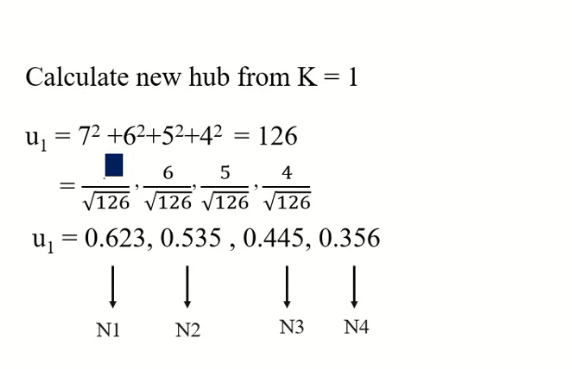


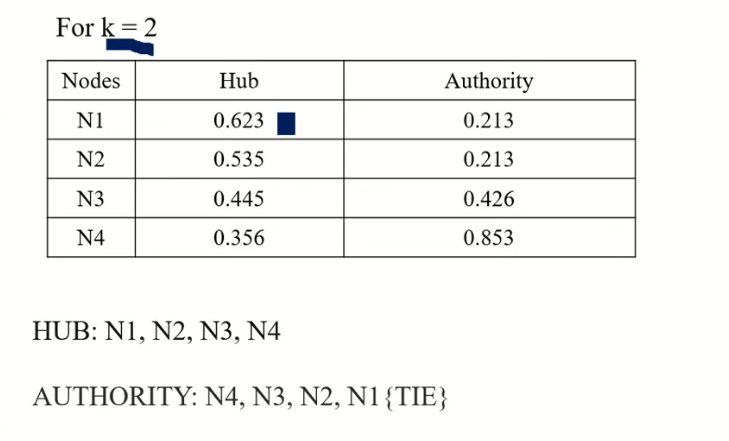


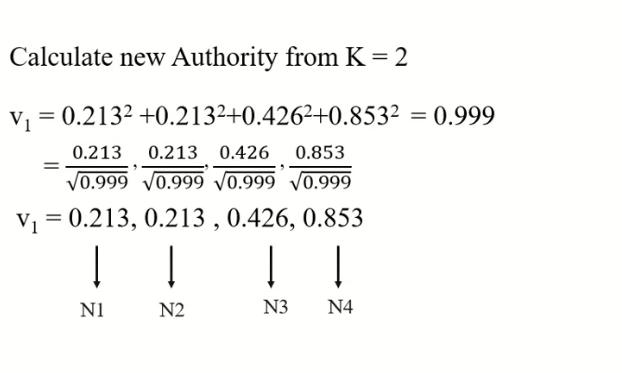












**Community Discovery**

Community Discovery Intuitively, a community is simply a group of entities (e.g., people or organizations) that shares a common interest or is involved in an activity or event. we showed that the HITS algorithm can be used to find communities. The communities are represented by dense bipartite sub-graphs. We now describe several other community finding algorithms. Apart from the Web, communities also exist in emails and text documents. This section describes two community finding algorithms for the Web, one community finding algorithm for emails, and one community finding algorithm for text documents.

There are many reasons for discovering communities.

1. Communities provide valuable and possibly the most reliable, timely, and up-to-date information resources for a user interested in them.

2. They represent the sociology of the Web: studying them gives insights into the evolution of the Web.

3. They enable target advertising at a very precise level.

Problem Definition Definition (community):

Given a finite set of entities S = {s1, s2, …, sn} of the same type, a community is a pair C = (T, G), where T is the community theme and G ⊆ S is the set of all entities in S that shares the theme T. If si ∈ G, si is said to be a member of the community C.

Some remarks about this definition are in order:

• A theme defines a community. That is, given a theme T, the set of members of the community is uniquely determined. Thus, two communities are equal if they have the same theme.

• A theme can be defined arbitrarily. For example, it can be an event (e.g., a sport event or a scandal) or a concept (e.g., Web mining).

• An entity si in S can be in any number of communities. That is, communities may overlap, or multiple communities may share members.

• The entities in S are of the same type. For example, this definition does not allow people and organizations to be in the same community.

• By no means does this definition cover every aspect of communities in the real world. For example, it does not consider the temporal dimension of communities. Usually a community exists within a specific period of time. Similarly, an entity may belong to a community during some time periods.

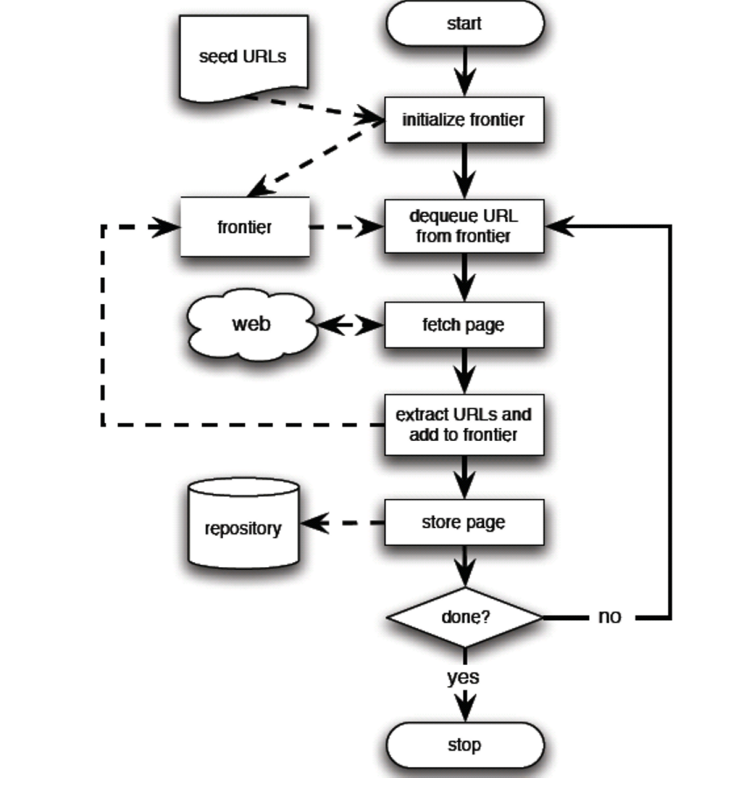
• This is a conceptual definition. In practice, different community mining algorithms have their own operational definitions which usually depend on how communities manifest themselves in the given data (which we will discuss shortly). Furthermore, the algorithms may not be able to discover all the members of a community or its precise theme.

**Web Crawling**

Web crawlers, also known as spiders or robots, are programs that automatically download Web pages. Since information on the Web is scattered among billions of pages served by millions of servers around the globe, users who browse the Web can follow hyperlinks to access information, virtually moving from one page to the next. A crawler can visit many sites to collect information that can be analyzed and mined in a central location, either online (as it is downloaded) or off-line (after it is stored). Were the Web a static collection of pages, we would have little long term use for crawling. Once all the pages are fetched and saved in a repository, we are done. However, the Web is a dynamic entity evolving at rapid rates. Hence there is a continuous need for crawlers to help applications stay current as pages and links are added, deleted, moved or modified.

is a continuous need for crawlers to help applications stay current as pages and links are added, deleted, moved or modified. There are many applications for Web crawlers. One is business intelligence, whereby organizations collect information about their competitors and potential collaborators. Another use is to monitor Web sites and pages of interest, so that a user or community can be notified when new information appears in certain places. There are also malicious applications of crawlers, for example, that harvest email addresses to be used by spammers or collect personal information to be used in phishing and other identity theft attacks. The most widespread use of crawlers is, however, in support of search engines. In fact, crawlers are the main consumers of Internet bandwidth. They collect pages for search engines to build their indexes. Well known search engines such as Google, Yahoo! and MSN run very efficient universal crawlers designed to gather all pages irrespective of their content. Other crawlers, sometimes called preferential crawlers, are more targeted. They attempt to download only pages of certain types or topics. This chapter introduces the main concepts, algorithms and data structures behind Web crawlers.

After discussing the implementation issues that all crawlers have to address, we describe different types of crawlers: universal, focused, and topical. We also discuss some of the ethical issues around crawlers. Finally, we peek at possible future uses of crawlers in support of alternative models where crawling and searching activities are distributed among a large community of users connected by a dynamic and adaptive peer network.



**Universal crawlers**

Universal crawlers, also known as general-purpose crawlers or web spiders, are web crawlers that are designed to traverse the entire World Wide Web and collect data from any publicly accessible web page.

Universal crawlers operate by starting from a set of seed URLs and then following links from those pages to discover other pages. They typically use techniques such as link analysis, keyword analysis, and content analysis to identify and index pages.

Universal crawlers are used by search engines to index web pages for search results, as well as by researchers and data scientists for web data mining and analysis. They are useful for collecting data on a wide range of topics and can provide a comprehensive view of the web.

However, universal crawlers face several challenges, including the large and constantly changing size of the web, the need to manage duplicate content and spam pages, and the need to prioritize pages based on relevance and importance. As a result, search engines and other organizations that use universal crawlers invest significant resources in optimizing their crawlers to provide high-quality and relevant results.

**A topical crawler**

A topical crawler, also known as a focused crawler, is a web crawler that is designed to search and index web pages based on a specific topic or set of topics. Unlike general-purpose crawlers that search and index web pages on all topics, a topical crawler is focused on a particular topic or domain, which allows it to provide more relevant results.

Topical crawlers use various techniques to identify and index pages related to a specific topic, such as keyword analysis, link analysis, and content analysis. They may also use machine learning algorithms to improve their ability to identify relevant pages and filter out irrelevant pages.

Topical crawlers are useful for applications such as information retrieval, search engine optimization, and web data mining, where specific information on a particular topic is needed. They can also be used to monitor changes in a particular domain or topic area over time.

**Focused crawlers**

Focused crawlers, also known as topical crawlers or selective crawlers, are web crawlers that are designed to gather and index information from specific parts of the web, such as a particular website or a particular topic.

Unlike general-purpose crawlers that crawl the entire web, focused crawlers target a specific domain or set of domains. They are especially useful when a researcher or organization wants to collect data from a limited set of web pages or websites.

Focused crawlers typically use one or more seed URLs to start their crawl and then follow links from those URLs to identify other pages related to the target domain or topic. They may also use techniques such as keyword analysis and content analysis to identify pages related to the target topic.

Focused crawlers are commonly used for web research, data mining, and other applications where targeted data collection is required. They can be used to collect data on a wide range of topics, from scientific research to e-commerce product information. However, they require more manual configuration and maintenance than general-purpose crawlers, as they need to be updated and configured with new seed URLs as the target domain or topic changes over time.

**Evaluation**

Evaluation in crawling refers to the process of measuring the performance of a web crawler, which includes assessing the quality and relevance of the web pages that the crawler collects and indexes.

There are several key metrics used to evaluate the effectiveness of a web crawler, including:

1. Coverage: This refers to the percentage of pages on the web that the crawler is able to discover and crawl. High coverage is desirable, as it means that the crawler is indexing a large number of pages relevant to the target domain or topic.
2. Precision: This refers to the percentage of pages that the crawler indexes that are relevant to the target domain or topic. High precision is desirable, as it means that the crawler is accurately identifying and indexing relevant pages.
3. Recall: This refers to the percentage of relevant pages on the web that the crawler is able to identify and index. High recall is desirable, as it means that the crawler is not missing a significant number of relevant pages.
4. Efficiency: This refers to the speed and resource efficiency of the crawler, including factors such as the time required to crawl a given number of pages and the resources required to run the crawler.

Evaluation in crawling is important to ensure that the crawler is performing optimally and providing high-quality results. It can also help identify areas for improvement, such as tuning the crawler's parameters, refining the target domain or topic, or improving the crawler's ability to identify relevant pages.

**Crawler ethics and conflicts**

Crawler ethics and conflicts refer to the ethical considerations and conflicts of interest that arise when collecting data using web crawlers.

One of the main ethical considerations is the privacy of individuals whose data may be collected by the crawler. This includes ensuring that the crawler does not collect personal information without consent and that the collected data is used in compliance with applicable privacy laws and regulations.

Another ethical consideration is the impact of crawling on the web and web servers. Crawlers can put a strain on web servers, causing them to slow down or crash. It is important for crawlers to operate within acceptable limits and to respect websites' terms of service and robots.txt files.

Conflicts of interest can also arise when using crawlers. For example, search engines may prioritize their own content over competitors' content, or companies may use crawlers to gain an advantage over competitors.

To address these ethical considerations and conflicts, organizations that use crawlers should have clear policies and guidelines in place for data collection, use, and storage. They should also work closely with website owners and webmasters to ensure that crawlers are operating within acceptable limits and are not causing harm.

In addition, organizations should consider using ethical frameworks, such as the ACM Code of Ethics, to guide their use of crawlers and ensure that they are operating in a responsible and ethical manner.