**UNIT I INTRODUCTION**

**Introduction to Semantic Web: Limitations of current Web - Development of Semantic Web - Emergence of the Social Web - Social Network analysis: Development of Social Network Analysis - Key concepts and measures in network analysis - Electronic sources for network analysis: Electronic discussion networks, Blogs and online communities - Web-based networks - Applications of Social Network Analysis.**

**1.1. Introduction to Semantic Web:**

* The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the application of advanced knowledge technologies to the web and distributed systems.
* The vision of the semantic web is that of a world wide distributed architecture where data and services easily interoperate.

**1.1.1. Limitations of current web:**

* The presence of huge amount of resources on the Web thus poses a serious problem of accurate search. This is mainly because today's Web is a human-readable Web where information cannot be easily processed by machine.
* Highly sophisticated, efficient keyword based search engines that have evolved today have not been able to bridge this gap.
* A search engine is a document retrieval system designed to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.
* Regardless of the underlying architecture, users specify keywords that match words in huge search engine databases, producing a ranked list of URLs and snippets of Web-pages in which the keywords matched.
* Although such technologies are mostly used, users are still often faced with the daunting task of shifting through multiple pages of results, many of which are irrelevant.
* The use of ontologies to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web.
* One of the biggest problems we nowadays face in the information society is information overload, a problem which is boosted by the huge size of the WWW. The Web has given us access to millions of resources, irrespective of their physical location and language.
* In order to deal with this sheer amount of information, new business models on the web have seen the light, such as commercial search engines. With the expected continuous growth of the WWW, we expect search engines will have a hard time maintaining the quality of retrieval results. Moreover, they only access static content, and ignore the dynamic part of the web. It is our vision that the technology of current generation of search engines has it’s limits. To be able deal with the continuous growth of the WWW (in size its languages and formats), we need to exploit other information. So here the Semantic Web help us.
* The current Web is based on HTML, which specifies how to layout a web page for human readers. HTML as such cannot be exploited by information retrieval techniques to improve results, which has thus to rely on the words that form the content of the page; hence it is restricted to keywords. Search engines are thus programmed in such a way that the first page shows a diversity of the most relevant links related to the keyword.
* The current Web has its limitations when it comes to :

1. finding relevant information

2. extracting relevant information

3. combining and reusing information

* Finding information on the current Web is based on keyword search. Keyword search has a limited recall and precision due to :

(a) Synonyms : e.g. Searching information about “Cars” will ignore Web pages that contain the word “Automobiles” even though the information on these pages could be relevant.

(b) Homonyms : e.g. Searching information about “Jaguar” will bring up pages containing information about both “Jaguar” (the car brand) and “Jaguar” (the animal) even though the user is interested only in one of them.

Keyword search has a limited recall and precision due also to :

1. Spelling variants : e.g. “organize” in American English vs. “organise” in British English

2. Spelling mistakes

3. Multiple languages : i.e. information about same topics in published on the Web on different languages (English, German, Italian,...)

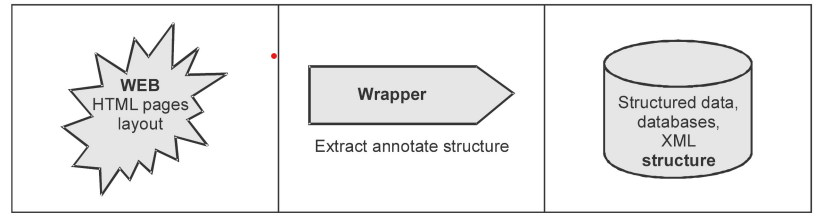
• Current search engines provide no means to specify the relation between a resource and a term : e.g. sell/buy.

• One-fit-all automatic solution for extracting information from web pages is not possible due to different formats, different syntaxes. Even from a single web page is difficult to extract the relevant information.

• Extracting information from current web sites can be done using wrappers.

• The actual extraction of information from web sites is specified using standards such as XSL Transformation.

• Extracted information can be stored as structured data in XML format or databases. However, using wrappers do not really scale because the actual extraction of information depends again on the web site format and layout.



• The actual extraction of information from web sites is specified using standards such as XSL Transformation.

• Extracted information can be stored as structured data in XML format or databases. However, using wrappers do not really scale because the actual extraction of information depends again on the web site format and layout.

**Knowledge gap**

* The knowledge gap is due to the lack of some kind of background knowledge that only the human possesses. The background knowledge is often completely missing from the context of the Web page and thus our computers do not even stand a fair chance by working on the basis of the web page alone.
* Semantic web is being to be developed to overcome the following problems for current web.

1. The web content lacks a proper structure regarding the representation of

information.

2. Ambiguity of information resulting from poor interconnection of

information.

3. Automatic information transfer is lacking.

4. Usability to deal with enormous number of users and content ensuring trust

at all levels.

5. Incapability of machines to understand the provided information due to lack

of a universal format.

**1.1.2. Development of Semantic Web:**

* The Semantic Web is an extension of the current web in which information is given welldefined meaning, better enabling computers and people to work in co-operation.
* It is a next generation of the WWW. Information has machine-processable and machineunderstandable semantics.
* Not a separate Web but an augmentation of the current one. The backbone of Semantic Web are ontologies.
* Semantic Web technology provides a basis for information sharing and performing some functions with generic software, but doesn't solve all application problems. The some of the areas that must be addressed by an application using Semantic Web technology.
* The main obstacle to provide better support to Web users is that, at present , the meaning of Web content is not machine accessible.
* Although there are tools to retrieve texts, but when it comes to interpreting sentence and extracting useful information for the user, the capabilities of current software are still very limited.
* In the Semantic Web, pages not only store content as a set of unrelated words in a document, but also code their meaning and structure.
* Semantic Web is set to become the future because it makes the understanding between humans and machines easy. Semantic web Design methodologies use ontology languages such as RDF, OWL to represent information internally.
* The Semantic Web has been actively promoted since by the World Wide Web Consortium, the organization that is chiefly responsible for setting technical standards on the Web.
* The core technology of the Semantic Web, logic-based languages for knowledge representation and reasoning have been developed in the research field of Artificial Intelligence.
* As the potential for connecting information sources on a Web-scale emerged, the languages that have been used in the past to describe the content of the knowledge bases of standalone expert systems have been adapted to the open, distributed environment of the Web.
* Since the exchange of knowledge in standard languages is crucial for the interoperability of tools and services on the Semantic Web, these languages have been standardized by the W3C as a layered set of languages.
* Research to develop languages that, on the one hand, allows human users to describe the meaning of words and, on the other hand, a computer to process these descriptions, began in the late 1980s, and led to what we now call ontology languages, in particular OWL.
* Researchers from the School of Computer Science played a pivotal role in this international effort. Specifically, they played central roles designing these languages based on logic, namely so-called 'Description Logics', and demonstrated that they are suitable by developing powerful tools to process and engineer ontologies in these languages.
* Tools for creating, storing and reasoning with ontologies have been primarily developed by university-affiliated technology startups and at research labs of large corporations.
* Over time the focus of Semantic Web research has been significantly extended to other topics. In addition to knowledge representation and reasoning topics from related communities, in particular from Databases, Data Mining, Information Retrieval and Computational Linguistics.

1**. Language Standards and - extensions :** The development of standardized knowledge representation languages was the starting point of Semantic Web research. Languages like DAML, OWL and RDF, but also representation languages for services such as DAML-S, OWL-S and WSMO were developed and various extensions were proposed, only some of which actually made it into the official language standard. Further, researchers discussed the use of other existing languages like XML as a basis for the Semantic Web.

**2. Logic and Reasoning :** Most of the language standards proposed for the Semantic Web are based on some for-mal logic. Thus extending existing logics to completely cover the respective standards as well as the development of scalable and efficient reasoning methods have been in the focus of research form the beginning

**3. Ontologies and Modelling :** The existence of language standards is necessary for Semantic Web applications, but it does not enable people to build the right models.

**4. Linked Data :** As a reaction, linked data has been proposed as a bottom-up approach, where data is converted into Semantic Web standards with minimal ontological commitment, published and linked to other data sources.

**1.1.3. Benefits of the Semantic Web:**

* Consistent mechanisms to model information from simple vocabularies to complex ontologies.
* A formal model approach ensures information reasoning outcomes.
* Data linking opportunities aimed at supporting better user experiences, and hence, improved business outcomes.
* A groundswell of activity in the development of open-source tools to exploit Semantic Web technologies and information.
* Standardised by the W3C indicating global consensus and open royalty-free specifications.

**1.2. Emergence of the Social Web:**

* The Web was a read-only medium for a majority of users. Upto 1990, web was combination of a telephone book and yellow pages. Some user was knows about hyperlinks.
* When web 2.0 was invented by Tim O’Reilly, attitude towards the web was changed.
* Web 2.0 tools allow libraries to enter into a genuine conversation with their users. Libraries are able to seek out and receive patron feedback and respond directly.
* In 2003, noticeable shift in how people and businesses were using the web and developing web-based applications.
* Tim O'Reilly said that 'Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as a platform, and an attempt to understand the rules for success on that new platform”.
* Many Web 2.0 companies are built almost entirely on user-generated content and harnessing collective intelligence. Google, MySpace, Flickr, YouTube and Wikipedia, users create the content, while the sites provide the platforms.
* The user is not only contributing content and developing open source software, but directing how media is delivered, and deciding which news and information outlets you trust.
* At that stage we thought the Web 2.0 stack was fairly empty, but since those days the extent that people collaborate, communication, and the range of tools and technologies have rapidly changed.
* Editing blogs and wikis did not require any knowledge of HTML any more. Blogs and wilds allowed individuals and groups to claim their personal space on the Web and fill it with content at relative ease.
* The first online social networks entered the field at the same time as blogging and wikis started to take off. Web 2.0 is the network as platform, spanning all connected devices.
* Web 2.0 applications are those that make the most of the intrinsic advantages of that platform. It delivers software as a continually-updated service that gets better the more people use it.
* Consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others.

**Web 3.0:**

* Web 3.0, or the Semantic Web, is the web era we are currently in, or perhaps the era we are currently creating. Web 3.0, with its use of semantics and artificial intelligence is meant to be a “smarter web”, one that knows what content you want to see and how you want to see it so that it saves you time and improves your life.
* Semantic Web is really the participatory web, which today includes “Classics” such as YouTube, MySpace, eBay, Second Life, Blogger, RapidShare, Facebook and so forth.
* Web 2.0 is that users are willing to provide content as well as metadata. This may take the form articles and facts organized in tables and categories in Wikipedia, photos organized in sets and according to tags in Flickr or structured information embedded into homepages and blog postings using micro-formats.
* A major disadvantage associated with Web 2.0 is that the websites become vulnerable to abuse since, anyone can edit the content of a Web 2.0 site. It is possible for a person to purposely damage or destroy the content of a website.
* Web 2.0 also has to address the issues of privacy. Take the example of YouTube. It allows any person to upload a video. But what if the video recording was done without the knowledge of the person who is being shown in the video? Thus, many experts believe that Web 2.0 might put the privacy of a person at stake.
* The basic idea of web 3.0 is to define structure data and link them in order to more effective discovery, automation, integration, and reuse across various applications. It is able to improve data management, support accessibility of mobile internet, simulate creativity and innovation, encourage factor of globalization phenomena, enhance customers' satisfaction and help to organize collaboration in social web.
* Web 3.0 supports world wide database and web oriented architecture which in earlier stage was described as a web of document.

**1.3. Social Network analysis:**

* Social Network Analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The term “social network” has been introduced by Barnes in 1954.
* SNA is the study of social relations among a set of actors. The methods of data collection in network analysis are aimed at collecting relational data in a reliable manner. Data collection is typically carried out using standard questionnaires and observation techniques that aim to ensure the correctness and completeness of network data.
* Social network analysis is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes.
* The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.
* The advantage of social network analysis is that, unlike many other methods, it focuses on interaction. Network analysis allows us to examine how the configuration of networks influences how individuals and groups, organizations, or systems function.
* Features of social network analysis : Structural intuition, systematic relational data, graphic representation and mathematical or computational models.

**Principles of Social Network Analysis:**

1. Actors and their actions are viewed as interdependent rather than independent, autonomous units.

2. Relational ties (linkages) between actors are channels for transfer or “flow” of resources (either material or nonmaterial).

3. Network models focusing on individuals view the network structure environment as providing opportunities for or constraints on individual action.

4. Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors.

**Social Network Analysis:**

1. Refers to the set of actors and the ties among them

2. Views on characteristics of the social units arising out of structural or relational processes or focuses on properties of the relational system themselves

3. Inclusion of concepts and information on relationships among units in a study

4. The task is to understand properties of the social (economic or political) structural environment, and

5. How these structural properties influence observed characteristics and associations among characteristics.

6. Relational ties among actors are primary and attributes of actors are secondary

7. Each individual has ties to other individuals, each of whom in turn is tied to a few, some, or many others, and so on

**Fundamental Concepts in Network Analysis:**

* Following terminology is used in social network analysis.

1. actor 2. relational tie 3. Dyad 4. triad 5. Subgroup 6. Group 7. relation

* **Actor :** Actor is discrete individual, corporate, or collective social units. Examples: people in a group, departments within in a corporation, public service agency in a city, nationstates in the world system.
* **Relational tie :** Actors are linked to another by social ties. A tie establishes a linkage between a pair of actors.
* **Dyad :** It is a tie between two actors and consists of a pair of actors and the tie(s) between them.
* **Triad :** Triples of actors and associated ties. A subset of three actors and the tie(s) among them.
* **Subgroup** of actors is defined as any subset of actors, and all ties among them.
* **Group :** Group is the collection of all actors on which ties are to be measured.
* **Relation :** It is the collection of ties of a specific kind among members of a group. Example : the set of friendship among pairs of children in a classroom
* Network can be categorized by the nature of the sets of actors and the properties of the ties among them. The number of modes in a network refers to the number of distinct kinds of social entities in the network.
* One-mode networks are a single set of actors. Two-mode networks are focus on two sets of actors, or one set of actors and one set of events.

**1.3.1. Development of Social Network Analysis:**

* A social network is a group of collaborating, and/or competing individuals or entities that are related to each other. It may be presented as a graph, or a multi-graphi each participant in the collaboration or competition is called an actor and depicted as a node in the graph theory.
* Valued relations between actors are depicted as links, or ties, either directed or undirected, between the corresponding nodes.
* Actors can be persons, organizations, or groups - any set of related entities. As such, SNA may be used on different levels, ranging from individuals, web pages, families, small groups, to large organizations, parties, and even to nations.
* In general, a social network consists of actors (e.g., persons, organizations) and some form of relation among them. The network structure is usually modeled as a graph, in which vertices represent actors, and edges represent ties, i.e., the existence of a relation between two actors.
* The vocabulary models and methods of network analysis also expand continuously through applications that require to handle ever more complex data sets.
* An example of this process are the advances in dealing with longitudinal data. New probabilistic models are capable of modelling the evolution of social networks and answering questions regarding the dynamics of communities. Formalizing an increasing set of concepts in terms of networks also contributes to both developing and testing theories in more theoretical branches of sociology.
* The purpose of social network analysis is to identify important actors, crucial links, roles, dense groups, and so on, in order to answer substantive questions about structure.
* Analysis methods available in visone are divided into four main categories according to the level or subject of interest: vertex, dyad, group, and network level .
* Available analysis methods include actor-level centrality indices, e.g. closeness, betweenness, and pagerank, cohesive subgroups like cliques, k-cliques, and k-clans, centrality and connectedness
* These levels break further down into measures of the same objective, e.g., connectedness or cohesiveness. Analysis methods are accessible using the analysis tab in the control area.

**1.4. Key concepts and measures in network analysis:**

* Social Network Analysis has developed a set of concepts and methods specific to the analysis of social networks.
* Several analytic tendencies distinguish social network analysis :

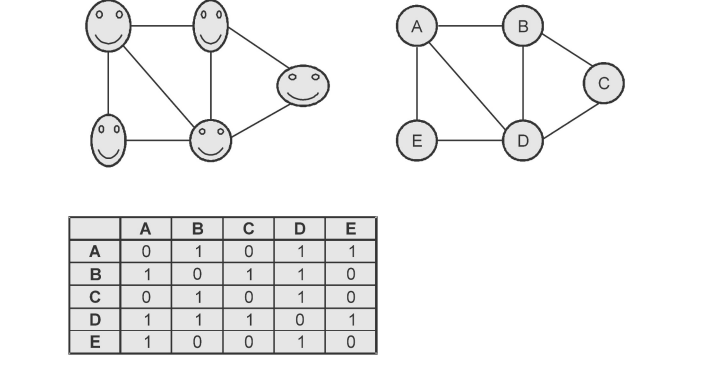
1. There is no assumption that groups are the building blocks of society : the approach is open to studying less-bounded social systems, from nonlocal communities to links among websites.

2. Rather than treating individuals (persons, organizations, states) as discrete units of analysis, it focuses on how the structure of ties affects individuals and their relationships.

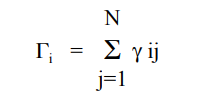
3. In contrast to analyses that assume that socialization into norms determines behavior, network analysis looks to see the extent to which the structure and composition of ties affect norms.

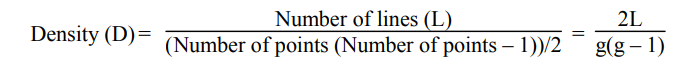
**1.4.1. Global Structure of Networks:**

* Social network can be represented as a graph G = (V, E) where V = The finite set of vertices E = Finite set of edges such
* The most network analysis methods work on an abstract, graph based representation of real world networks.



***Graph based representation of real world networks***

* When representing a network as a graph, all of the connections are pair-wise and hence represented by ties known as edges.
* Networks can be described using a mixture of local, global and intermediate-scale perspectives. Accordingly, one of the key uses of network theory is the identification of summary statistics for large networks in order to develop a framework for analyzing and comparing complex structures.
* SNA can produce maps like the one featured below and provide statistical measures of relationships between actors. In SNA maps, the nodes represent the different actors in the network and the lines represent the relationships between the various actors.
* The size of the node often represents the relative importance of that actor in the network and the thickness of the connecting line denotes the strength of the relationship.
* Clustering for a single vertex can be measured by the actual number of the edges between the neighbours of a vertex divided by the possible number of edges between the neighbours.
* When taken the average over all vertices, we get to the measure known as clustering coefficient. The clustering coefficient of a tree is zero, which is easy to see if we consider that there are no triangles of edges (triads) in the graph. In a tree, it would never be the case that our friends are friends with each other.
* The coordination degree measures the ability of the vertices in a graph to interchange information. There are several ways in which we can model this magnitude. One of the easiest is to consider the coordination degree to be exponentially related with the distance between the vertices.
* To define the total co-ordination degree of a vertex “i” in a graph as the sum of all the coordination degree between that particular vertex and the rest :
* 
* Where N is the order of the graph
* Graph density (D) is defined as the total number of observed lines in a graph divided by the total number of possible lines in the same graph. Density ranges from 0 to 1.



**Random Graphs with Arbitrary Degree Distributions:**

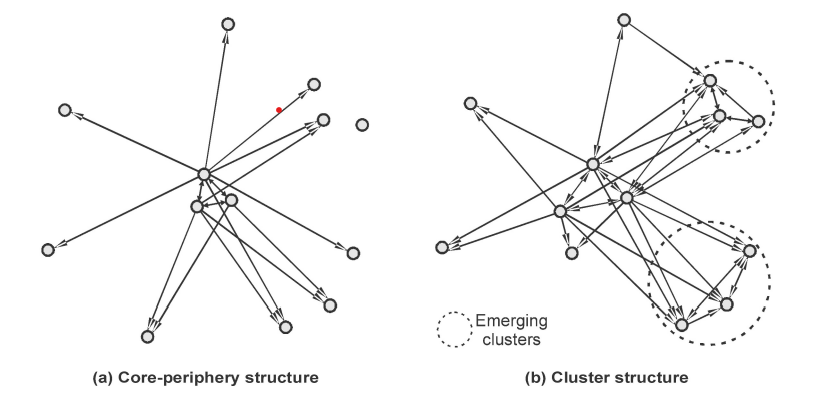
* A random graph is simple to define. One takes some number N of nodes or “vertices” and places connections or “edges” between them, such that each pair of vertices i, j has a connecting edge with independent probability p.
* Random graph can be generated by taking a set of vertices with no edges connection them. Subsequently, edges are added by picking pairs of nodes with equal probability.
* Consider a vertex in a random graph. It is connected with equal probability p with each of the N – 1 other vertices in the graph and hence the probability pk that it has degree exactly k is given by the binomial distribution :



* A large random graph has a Poisson degree distribution. This degree distribution makes the random graph a poor approximation to the real world networks.

**Macro-structure of social networks:**

* Network visualizations based on topographic or physical principles can be helpful in understanding the group structure of social networks and pinpoint hubs that naturally tend to gravitate toward the centre of the visualization.
* Clustering a graph into subgroups allows us to visualize the connectivity at a group level.
* Core-Periphery structure is one where nodes can be divided in two distinct subgroups : nodes in the core are densely connected with each other and the nodes on the periphery, while peripheral nodes are not connected with each other, only nodes in the core.
* By computing a network’s core-periphery structure, one attempts to determine which nodes are part of a densely connected core and which are part of a sparsely connected periphery.
* Core nodes should also be reasonably well-connected to peripheral nodes, but the latter are not well-connected to a core or to each other.
* Node belongs to a core if and only if it is well-connected both to other core nodes and to peripheral nodes. A core structure in a network is thus not merely densely connected but also tends to be ‘central’ to the network.
* From network theory has it defined the dual relationships between nodes in the network, so that if an agent has a feature no other, for example, if it is good then it is not bad, is a bipartition graph in which each element of a subset is additional to another concept indeed implies that binding of the n subgroups partitions make the whole graph. So CPS, involves dividing the nodes of the network into two groups.



This figure shows core-periphery structure that would be perfect without the edge between nodes.

**Affiliation networks** contain information about the relationships between two sets of nodes : a set of subjects and a set of affiliations. An affiliation network can be formally represented as a bipartite graph, also known as a two-mode network.

• Affiliation networks are two mode networks that allow one to study the dual perspectives of the actors and the events. They look at collections or subsets of actors or subsets rather than ties between pairs of actors. Connections among members of one of the modes as based on linkages established through the second mode.

• An affiliation network is a network in which actors are joined together by common membership of groups or dubs of some kind.

• A distinctive feature of affiliation networks is duality i.e. events can be described as collections of individuals affiliated with them and actors can be described as collections of events with which they are affiliated.

Based on two-mode matrix data, affiliation networks consist of sets of relations connecting actors and events, rather than direct ties between pairs of actors as in one-mode data. Familiar affiliation networks include persons belonging to associations, social movement activists participating in protest events, firms creating strategic alliances, and nations signing treaties.

The representation of two-mode data should facilitate the visualization of three kinds of patterning : a) the actor-event structure b) the actor-actor structure c) the event-event structure

Many ways to represent affiliation networks :

1. Affiliation network matrix

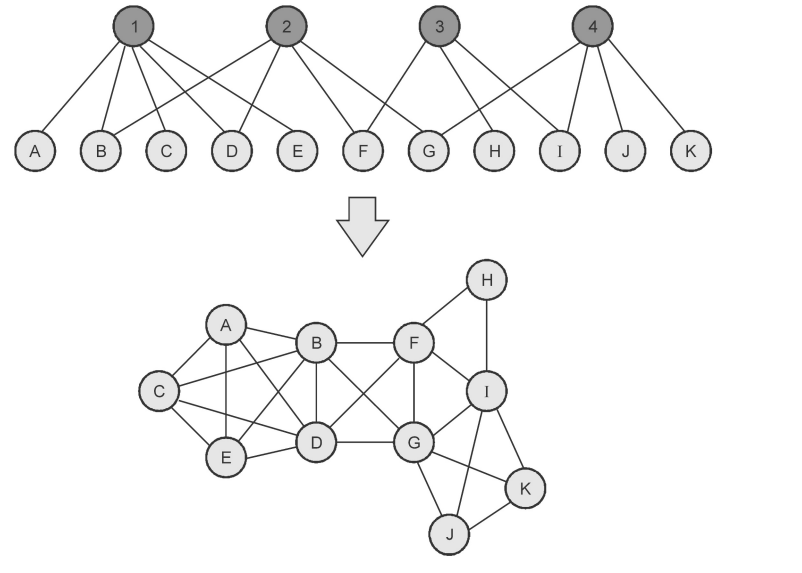
2. Bipartite graph or Sociomatrix

3. Hypergraph

4. Simplicial Complex

**Bipartite Graph:**

Nodes are partitions into two subsets and all lines are between pairs of nodes belonging to different subsets. The following figure shows bipartite network. As there are g actors and h events, there are g + h nodes.



Bipartite Graph

* “The lines on the graph represent the relation “is affiliated with” from the perspective of the actor and the relation ''has as a member” from the perspective of the event.
* No two actors are adjacent and no two events are adjacent. If pairs of actors are reachable, it is only via paths containing one or more events. Similarly, if pairs of events are reachable, it is only via paths containing one or more actors.

**Advantages:**

1. They highlight the connectivity in the network, as well as the indirect chains of connection.

2. Data is not lost and we always know which individuals attended which events.

**Disadvantage:**

1. They can be unwieldy when used to depict larger affiliation networks.

**Benefits of Affiliations Networks:**

1. Affiliations of actors with events provide a direct linkage between actors through memberships in events, or between events through common memberships.
2. Affiliations provide conditions that facilitate the formation of pairwise ties between actors.
3. Affiliations enable us to model the relationships between actors and events as a whole system.

**1.5. Electronic sources for network analysis:**

* Collecting social network data used to be a tedious, labor-intensive process. In fact, several notable dissertations came out of the researcher’s being at the right place and the right time to be able to observe a social conflagration and gather data on it.
* Social network data collection is, by nature, more invasive and harder to anonymize; survey instruments had to be approved by Institutional Review Boards (IRBs), and administration of the surveys was tedious manual labor.
* Some of key challenges in this kind of data collection are :

1. Network boundaries are difficult to define.

2. People do not easily recall their network members, and need appropriate “prompts” to elicit them. In addition, networks are very large in general, and different social network members may have different importance depending on the phenomenon studied.

3. Information about the network members needs to balance detail and interviewee's burden.

* Most social network data collection can be divided into “whole” and “egocentric” networks. Whole network studies examine actors “that are regarded for analytical purposes as bounded social collectives”; actors in these studies are named in closed lists, usually predefined, and known a priori.
* Since these boundaries are very difficult to define in urban settings with large populations, whole network studies are unpractical, making egocentric data collection the only feasible method.
* Egocentric network studies concentrate in specific actors or egos and those who have relations with them, called alters. That is, from the participant's perspective, egocentric networks constitute a “network of me” or a network of actors with whom the participant has some relationship.
* Egocentric network data is thus composed by two levels:

i) an ego-network level, constituted by the ego's characteristics and overall network features; and

ii) an ego-alter level, constituted by the characteristics of each alter and alter-ego ties.

**1.5.1. Electronic discussion networks:**

* The study of the email network useful in identifying leadership roles within the organization and finding formal as well as informal communities.
* Wu, Huberman, Adamic and Tyler use this data set to verify a formal model of information flow in social networks based on epidemic models.
* Adamic and Adar revisits one of the oldest problems of network research, namely the question of local search : how do people find short paths in social networks based on only local information about their immediate contacts ?
* Even the Huge and versatility of data, the studies of electronic communication networks based on email data are limited by privacy reasons. Public forums and mailing lists can be analyzed without similar fashion.
* Group communication and collective decision taking in various settings are traditionally studied using much more limited written information such as transcripts and records of attendance and voting.

**1.5.2. Blogs and online communities:**

* A blog was selected as it facilitated and encouraged rich and deep reflection since the participants had to put their thoughts into writing and they had the time to reflect on what they were really experiencing.
* Blogs, like diaries, are continuous. Once the blog went live, it was available to the participants throughout the six-month data collection period.
* Blogs can also be considered only minimally intrusive on participant's lives since users can access the blog whenever they wish - just as with traditional diaries. Blogs also facilitate the collection of data across several geographical locations simultaneously.
* Like diaries, blogs are multimodal. They facilitate different kinds of expression. In this way, the blog honours participants' voices and the individual ways in which they may find their voices.
* On blogs, users could express themselves using several forms of text including, but not limited to, narratives, comments and poetry.
* Further, the medium allows users to upload or post links to pictures, art, video and music which are meaningful to the participants in some way.
* Fundamentally, blogs are interactive. While this is a major departure from traditional diaries, it was thought that the interactive nature of the blog would help to hold participant interest and to keep data collection progressing where traditional diaries had shown to become monotonous.
* In addition, it was felt that the interactive feature of the blog would give the participants something in return for their assistance. It would give them the opportunity to meet and interact with other Trinidadian students in the UK and learn about others' experiences while being able to share their own thoughts, feelings and experiences and receive feedback.
* A consistent and significant problem that many researchers face when conducting research online is the anonymity of the participants. This is particularly problematic when acquiring informed consent from people who the researchers do not know or cannot see.
* However, for researchers using the internet as a research tool, and not as the site of the study, the anonymity that the internet provides can be perceived as strength rather than a limitation.
* The anonymity provided by the internet has also been shown in some studies to reduce anxieties about feeling judged and can increase self-disclosure motivating deeper introspection and reflection.
* Over time, a blog can also encourage a community atmosphere among group members, increasing comfort levels and making it easier for participants to self-disclose.
* Blogs were also accessible for the research population and particularly suited to them. As university students, the participants have unlimited access to the internet.
* Further, computers are a necessary component in students' lives. It is where they conduct research, write papers, access their university email, manage everyday student administrative needs, contact lecturers and get involved in classes. The method seemed both relevant and accessible to the research population. The procedure for using a blog - as a research participant - is similar to logging into any general internet service, creating, and then sending an email, and could be easily taught to interested participants.

**1.5.3. Web-based networks:**

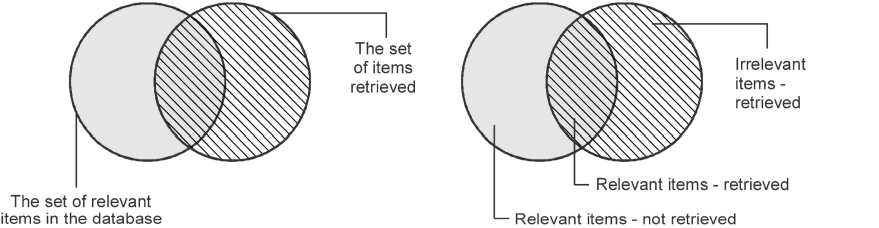
* Content of Web pages is the most inexhaustible source of information for social network analysis. This content is not only vast, diverse and free to access but also in many cases more up to date than any specialized database.
* Features of web pages for extracting social relations are links and co-occurrences.
* Co-occurrence, which is also referred to as “implied links.” Co-occurrence is the relationship between similar words on a page and their proximity to brands and also links.
* In fact, Google filed the co-occurrence patent on June 30, 2011 to refine the search results that identify the most significant keyword and create a relationships between the related terms. Co-occurrence is a factor in ranking web pages for specific queries.
* Co-occurrences of names in web pages can also be taken as evidence of relationships and are a more frequent phenomenon. On the other hand, extracting relationships based on cooccurrence of the names of individuals or institutions requires web mining as names are typically embedded in the natural text of web pages.
* The link prediction problem is also related to the problem of inferring missing links from an observed network : in a number of domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist.
* In response to a query, an IR system searches its document collection and returns a ordered list of responses. It is called the retrieved set or ranked list. The system employs a search strategy or algorithm and measure the quality of a ranked list.
* A better search strategy yields a better ranked list and better ranked lists help the user fill their information need.
* Precision and recall are the basic measures used in evaluating search strategies.

As shown in the first two figures, these measures assume :

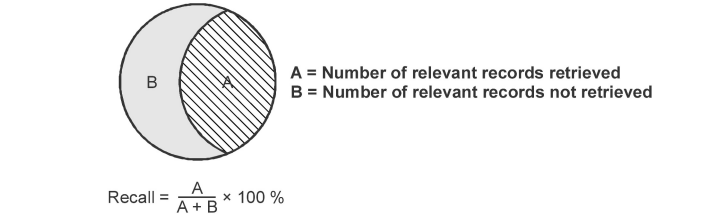
1. There is a set of records in the database which is relevant to the search topic.

2. Records are assumed to be either relevant or irrelevant.

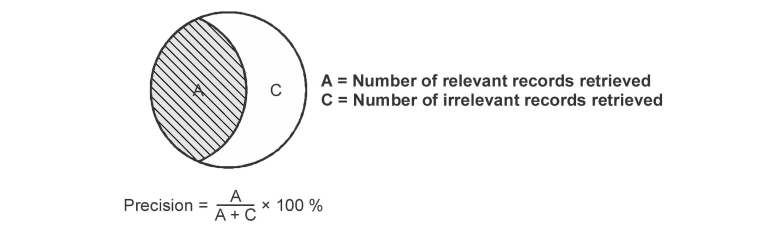
3. The actual retrieval set may not perfectly match the set of relevant records.



* Recall is the ratio of the number relevant records retrieved to the total number of relevant records in the database. It is usually expressed as percentage.

****

* Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

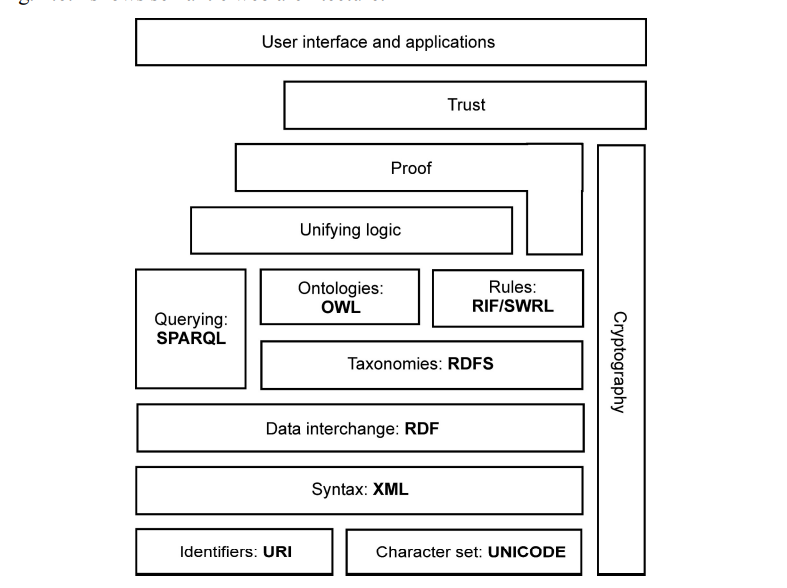
****

* As recall increases, the precision decreases and recall decreases the precision increases.
* The average precision method is more sophisticated in that it takes into account the order in which the search engine returns document for a person : it assumes that names of other persons that occur closer to the top of the list represent more important contacts than names that occur in pages at the bottom of the list. The method is also more scalable as it requires only downloading the list of top ranking pages once for each author.

**1.6. Applications of Social Network Analysis:**

* Social network analysis (SNA) is an important and valuable tool for knowledge extraction from massive and un-structured data. Social network provides a powerful abstraction of the structure and dynamics of diverse kinds of inter-personal connection and interaction.
* Facebook is a social networking service and website that connects people with other people, and share data between people. A user can create a personal profile, add other users as friends, exchange data, create and join common interest communities.
* Twitter is a social networking and microblogging service. The users of Twitter can exchange text-based posts called tweets. A tweet is a maximum 140 characters long but can be augmented by pictures or audio recording. The main concept of Twitter was to build a social network formed by friends and followers. Friends are people who you follow, followers are those who follow you.
* The role of social networks in labor markets deserves attention for at least two reasons : first, because of the central role networks play in disseminating information about job openings they place a critical role in determining whether labor markets function efficiently; and second, because network structure ends up having implications for things like human capital investment as well as inequality.
* Social network analysis (SNA) primarily focuses on applying analytic techniques to the relationships between individuals and groups, and investigating how those relationships can be used to infer additional information about the individuals and groups.
* SNA is used in a variety of domains. For example, business consultants use SNA to identify the effective relationships between workers that enable work to get done; these relationships often differ from connections seen in an organizational chart.
* Law enforcement personnel have used social networks to analyze terrorist networks and criminal networks. The capture of Saddam Hussein was facilitated by social network analysis : military officials constructed a network containing Hussein's tribal and family links, allowing them to focus on individuals who had close ties to Hussein.

**1.6.1. Generic Architecture of Semantic Web Applications:**

****

***Generic architecture of semantic web applications***

* The first layer, URI and Unicode, follows the important features of the existing WWW. Unicode is a standard of encoding international character sets and it allows that all human languages can be used on the web using one standardized form.
* URI is a string of a standardized form that allows to uniquely identify resources (e.g., documents).
* A subset of URI is Uniform Resource Locator (URL), which contains access mechanism and a location of a document.
* Another subset of URI is URN that allows to identify a resource without implying its location and means of dereferencing it.
* The usage of URI is important for a distributed internet system as it provides understandable identification of all resources.
* An international variant to URI is Internationalized Resource Identifier (IRI) that allows usage of Unicode characters in identifier and for which a mapping to URI is defined. In the rest of this text, whenever URI is used, IRI can be used as well as a more general concept
* Extensible Markup Language (XML) layer with XML namespace and XML schema definitions makes sure that there is a common syntax used in the semantic web. XML is a general purpose markup language for documents containing structured information
* A core data representation format for semantic web is Resource Description Framework (RDF). RDF is a framework for representing information about resources in a graph form.
* More detailed ontologies can be created with Web Ontology Language OWL. The OWL is a language derived from description logics, and offers more constructs over RDFS.

**1.6.2. Advantages and Disadvantages of Social Media:**

**Advantages of Social Media:**

**1. Brand awareness :** Compelling and relevant content will grab the attention of potential customers and increase brand visibility

**2. Brand reputation :** You can respond instantly to industry developments and be seen as ‘thought leader’ or expert in your field. This can improve how your business is seen by your audience

3. Brand loyalty : You can build relationships with your customers through social media. This can help increase loyalty and advocacy

**4. Customer interaction :** You can deliver improved customer service and respond effectively to feedback. Positive feedback is public and can be persuasive to other potential customers. Negative feedback highlights areas where you can improve.

**5. Target audience :** Customers can find you through the social media platforms they use most. You can choose to maintain a presence on particular social networks that are in line with your target audience

**6. Website traffic :** Social content can boost traffic to your website. This can lead to increased online conversions such as sales and leads.

**7. Cost effective :** It can be much cheaper than traditional advertising and promotional activities. 8. Evaluation : It is easy to measure how much website traffic you receive from social media. You can set up tracking to determine how many sales are generated by paid social advertising.

**Disadvantages of Social Media:**

**1. Resources :** You will need to commit resources to managing your social media presence, responding to feedback and producing new content.

**2. Evaluation :** While it is easy to quantify the return-on-investment in terms of online sales generated by social media

**3. Advertising :** It’s difficult to know how social media effects sales in-store.

**4. Ineffective use :** Social media can be used ineffectively. For example, using the network to push for sales without engaging with customers, or failing to respond to negative feedback, it may damage your reputation.