SUBJECT CODE : CS8085

Strictly as per Revised Syllabus of

ANNA UNIVERSITY

Choice Based Credit System (CBCS)

Semester - VIII (CSE / IT)
Professional Elective IV

SOCIAL NETWORK ANALYSIS

Iresh A. Dhotre
M.E. (Information Technology)
Ex-Faculty, Sinhgad College of Engineering,
Pune.

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

Social Network Analysis

Subject Code : CS8085

Semester - VIII (CSE / IT) Professional Elective - IV

Published by :

TECHNICAL PUBLICATIONS
SINCE 1993 An Up-Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,
Pune - 411030, M.S. INDIA, Ph.: +91-020-24495496/97
Email : sales@technicalpublications.org Website : www.technicalpublications.org

Printer :
Yogiraj Printers & Binders
Sr.No. 10/1A,
Ghule Industrial Estate, Nanded Village Road,
Tal. - Haveli, Dist. - Pune - 411041.

ISBN 978-93-90450-29-9

9 789390 450299

AU 17

PREFACE

The importance of **Social Network Analysis** is well known in various engineering fields. Overwhelming response to my books on various subjects inspired me to write this book. The book is structured to cover the key aspects of the subject **Social Network Analysis.**

The book uses plain, lucid language to explain fundamentals of this subject. The book provides logical method of explaining various complicated concepts and stepwise methods to explain the important topics. Each chapter is well supported with necessary illustrations, practical examples and solved problems. All the chapters in the book are arranged in a proper sequence that permits each topic to build upon earlier studies. All care has been taken to make students comfortable in understanding the basic concepts of the subject.

Representative questions have been added at the end of each section to help the students in picking important points from that section.

The book not only covers the entire scope of the subject but explains the philosophy of the subject. This makes the understanding of this subject more clear and makes it more interesting. The book will be very useful not only to the students but also to the subject teachers. The students have to omit nothing and possibly have to cover nothing more.

I wish to express my profound thanks to all those who helped in making this book a reality. Much needed moral support and encouragement is provided on numerous occasions by my whole family. I wish to thank the **Publisher** and the entire team of **Technical Publications** who have taken immense pain to get this book in time with quality printing.

Any suggestion for the improvement of the book will be acknowledged and well appreciated.

Author

D.A. Dhotre

Dedicated to God

SYLLABUS

Social Network Analysis - CS8085

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.

UNIT II MODELLING, AGGREGATING AND KNOWLEDGE REPRESENTATION

Ontology and their role in the Semantic Web : Ontology-based knowledge Representation - Ontology languages for the Semantic Web: Resource Description Framework - Web Ontology Language - Modelling and aggregating social network data: State-of-the-art in network data representation - Ontological representation of social individuals - Ontological representation of social relationships - Aggregating and reasoning with social network data - Advanced representations.

UNIT III EXTRACTION AND MINING COMMUNITIES IN WEB SOCIAL NETWORKS

Extracting evolution of Web Community from a Series of Web Archive - Detecting communities in social networks - Definition of community - Evaluating communities - Methods for community detection and mining - Applications of community mining algorithms - Tools for detecting communities social network infrastructures and communities - Decentralized online social networks - Multi-Relational characterization of dynamic social network communities.

UNIT IV PREDICTING HUMAN BEHAVIOUR AND PRIVACY ISSUES

Understanding and predicting human behaviour for social communities - User data management - Inference and Distribution - Enabling new human experiences - Reality mining - Context - Awareness - Privacy in online social networks - Trust in online environment - Trust models based on subjective logic - Trust network analysis - Trust transitivity analysis - Combining trust and reputation - Trust derivation based on trust comparisons - Attack spectrum and countermeasures.

UNIT V VISUALIZATION AND APPLICATIONS OF SOCIAL NETWORKS

Graph theory - Centrality - Clustering - Node-Edge Diagrams - Matrix representation - Visualizing online social networks, Visualizing social networks with matrix-based representations - Matrix and Node-Link Diagrams - Hybrid representations - Applications - Cover networks - Community welfare - Collaboration networks - Co-Citation networks.

TABLE OF CONTENTS

❑❑❑

1 Introduction

Scope of the Syllabus

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

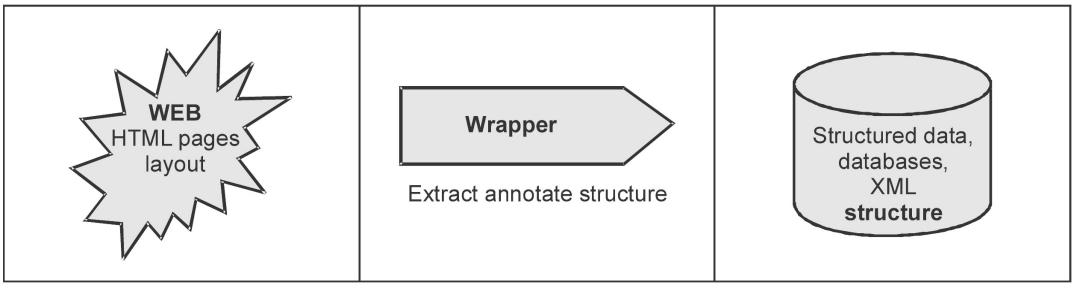
- The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the application of advanced knowledge technologies to the web and distributed systems.
- The vision of the 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.

➤ 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 well-defined meaning, better enabling computers and people to work in co-operation.
- It is a next generation of the WWW. Information has machine-processable and machine-understandable 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 stand-alone 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, nation-states 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.

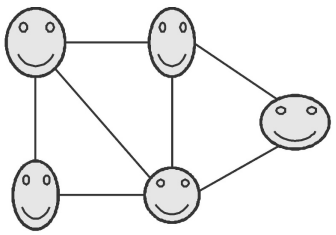
1.4.1 Global Structure of Networks

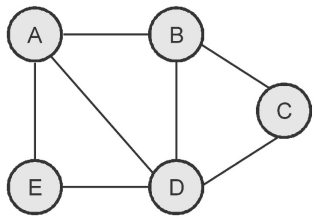
- Social network can be represented as a graph G = (V, E)

  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. It is shown in Fig. 1.4.1.





	A	B	C	D	E
A	0	1	0	1	1
B	1	0	1	1	0
C	0	1	0	1	0
D	1	1	1	0	1
E	1	0	0	1	0

Fig. 1.4.1 : 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 :

$$\Gamma_i = \sum_{j=1}^{N} \gamma_{ij}$$

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.

$$\text{Density (D)} = \frac{\text{Number of lines (L)}}{(\text{Number of points (Number of points} - 1))/2} = \frac{2L}{g(g-1)}$$

➤ 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 p_k that it has degree exactly k is given by the binomial distribution :

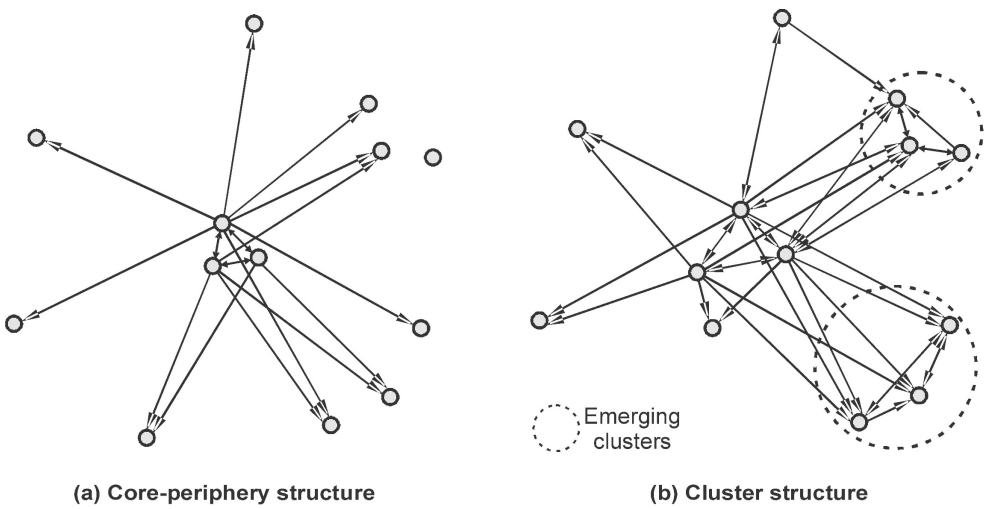
$$p_k = \begin{bmatrix} N-1 \\ k \end{bmatrix} p_k (1-p)^{N-1-k}$$

- 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.
- Fig. 1.4.2 shows core-periphery structure that would be perfect without the edge between nodes.



(a) Core-periphery structure (b) Cluster structure

Fig. 1.4.2

- **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. Fig. 1.4.3 shows bipartite network. As there are g actors and h events, there are g + h nodes.

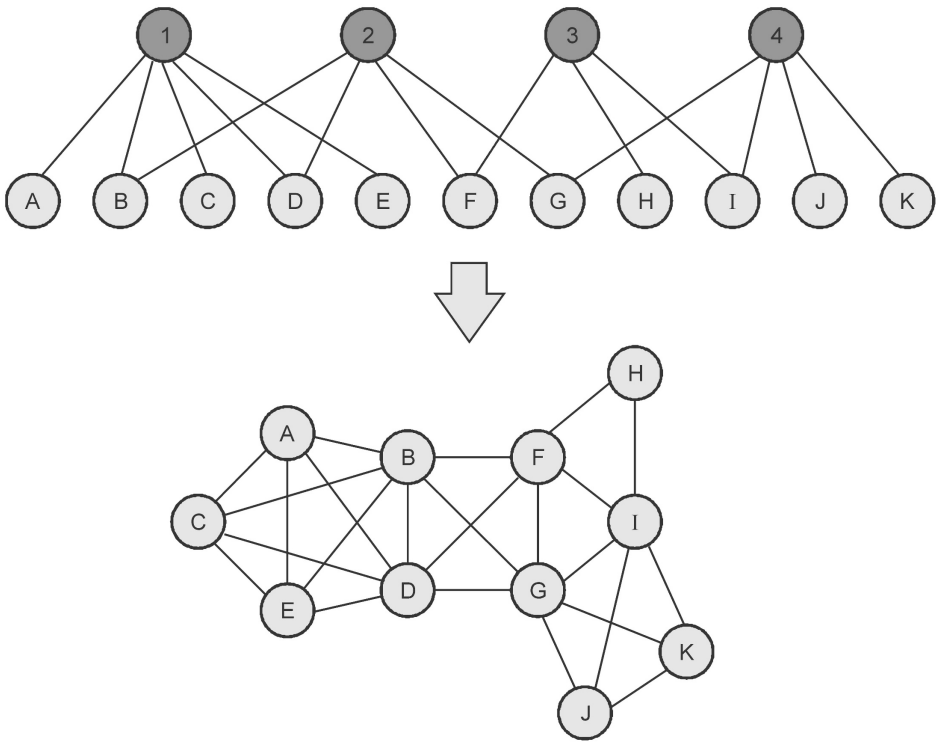


Fig. 1.4.3 : Bipartite graph

- "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 Network

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 pre-defined, 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

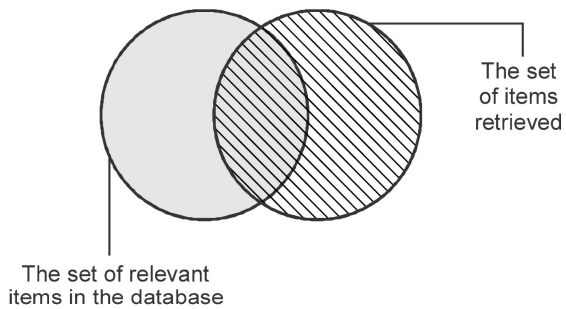
- A blog was selected as it facilitated and encouraged rich and deep reflection since the participants had to put their thoughts into writing and they had the time to reflect on what they were really experiencing.
- Blogs, like diaries, are continuous. Once the blog went live, it was available to the 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.

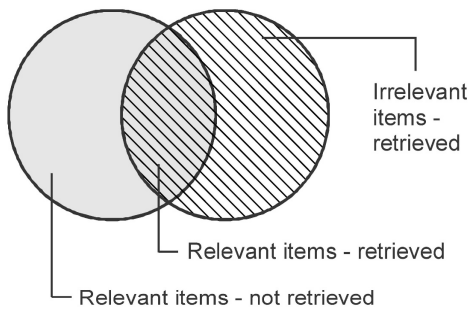
- Fundamentally, blogs are interactive. While this is a major departure from traditional diaries, it was thought that the interactive nature of the blog would help to hold participant interest and to keep data collection progressing where traditional diaries had shown to become monotonous.
- In addition, 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

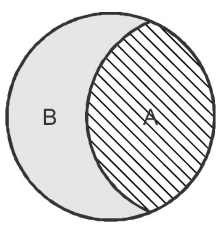
- Content of Web pages is the most inexhaustible source of information for social network analysis. This content is not only vast, diverse and free to access but also in many cases more up to date than any specialized database.
- Features of web pages for extracting social relations are links and co-occurrences.
- 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 co-occurrence 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.





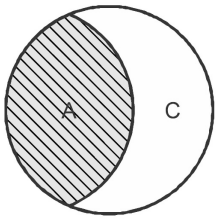
- Recall is the ratio of the number relevant records retrieved to the total number of relevant records in the database. It is usually expressed as percentage.



A = Number of relevant records retrieved
B = Number of relevant records not retrieved

$$\text{Recall} = \frac{A}{A + B} \times 100\ \%$$

- **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.



A = Number of relevant records retrieved
C = Number of irrelevant records retrieved

$$\text{Precision} = \frac{A}{A + C} \times 100\ \%$$

- 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

- Fig. 1.6.1 shows semantic web architecture.

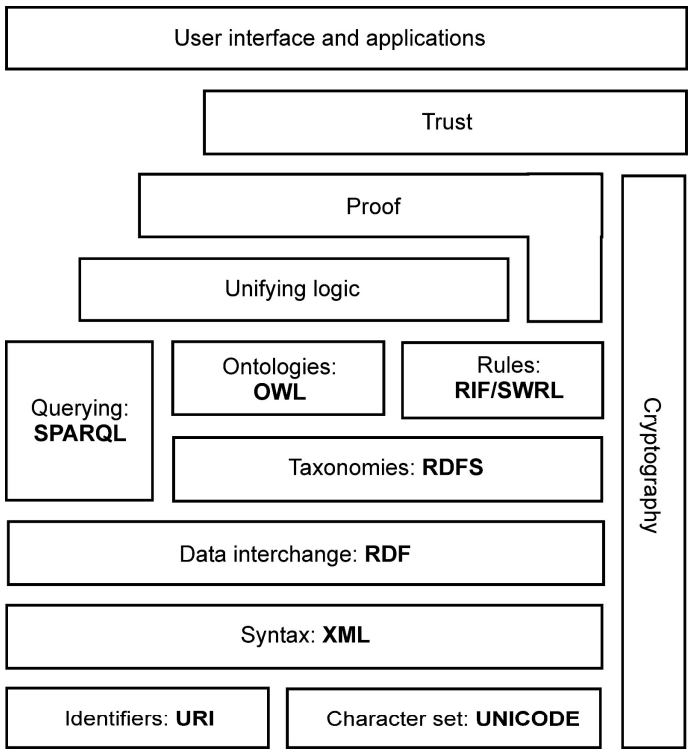


Fig. 1.6.1 : Generic architecture of semantic web applications

- 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.

1.7 Questions with Answers

1.7.1 Two Marks Question with Answer

Q. 1 What is semantic web ?

Ans. : The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation

Q. 2 List limitations of the current web search.

Ans. :

1. The Web search results are high recall, low precision.
2. Results are highly sensitive to vocabulary.
3. Results are single Web pages.
4. Most of the publishing contents are not structured to allow logical reasoning and query answering.

Q. 3 Define ontology.

Ans. : An ontology is a hierarchically structured set of terms to describe a domain that can be used as a skeletal foundation for a knowledge base.

Q. 4 What is a social network ?

Ans. : A social network is a group of collaborating, and/or competing individuals or entities that are related to each other. Social network is formally defined as a set of social actors, or nodes, members that are connected by one or more types of relations.

Q. 5 What is Social Network Analysis ?

Ans. : Social Network Analysis (SNA) is the study of social relations among a set of actors.

Q. 6 What is affiliation network ?

Ans. : Affiliation networks are **two mode networks** that allow one to study the dual perspectives of the actors and the events

Q. 7 Explain adjacency matrix.

Ans. : An adjacency matrix is a square matrix with one row and one column for each vertex in a network. The content of a cell in the matrix indicates the presence and possibly the sign or value of a tie from the vertex represented by the row to the vertex represented by the column

Q. 8 What is transitivity model ?

Ans. : The transitivity model applies to an unsigned directed network if it consists of cliques such that cliques within ranks are not related and cliques between ranks are related by null dyads or asymmetric dyads pointing towards the higher rank.

Q. 9 What is two mode network ?

Ans. : In a two-mode network, vertices are divided into two sets and vertices can only be related to vertices in the other set.

Q. 10 Define precision.

Ans. : 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.

1.7.2 Fill in the Blanks

Q. 1 Co-occurrence is also referred to as ---------- links.

Q. 2 Twitter is a social networking and ---------- service.

Q. 3 The goal of the Semantic Web is to make Internet data ----------.

Q. 4 Web 3.0 is also called as ---------- web.

Q. 5 Affiliation networks are -------- networks.

Q. 6 ---------- networks are bipartite networks denoting the membership of actors in groups.

Q. 7 HTML stands for ----------.

Q. 8 The idea of the Semantic Web is to apply advanced ---------- in order to fill the knowledge gap between human and machine

Q. 9 Information that is missing or hard to access for our machines can be made accessible using ---------.

Q. 10 The --------- --- gap is due to the lack of some kind of background knowledge that only the human possesses

Q. 11 --------- is based on the Music Genome Project, an attempt to create a vocabulary to describe characteristics of music from melody, harmony and rhythm.

Q. 12 ------- is the ratio of the number of relevant records retrieved to the total number of relevant records in the database

Q. 13 The Resource description framework was the first and is the fundamental technology of the -------- Web.

Q. 14 The --------- of a graph is quantitatively defined as the number of links divided by the number of vertices in a complete graph with the same number of nodes.

1.7.3 Multiple Choice Questions

Q. 1 Acronym for RDF is

(a) Resource Data Framework (b) Resource Description Framework

(c) Object Description Framework (d) Relationship Object Framework

Q. 2 The first step for semantic web is

(a) HTML (b) XML (c) RDF (d) WSMO

Q. 3 Online social networks represent ties between persons in online social networking platforms.

(a) ties (b) actor (c) dyad (d) triad

Q. 4 Affiliation networks are represented by

(a) hypergraph (b) Bipartite graph

(c) affiliation network matrix (d) All of these

Q. 5 Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection.

(a) Recall (b) Precision (c) relevant (d) none of these

Q. 6 Precision is the ratio of the number of ---------- retrieved to the total number of documents retrieved.

(a) Unrelevant documents (b) relevant documents

(c) relevant informations (d) None of these

Q. 7 XML stands for eXtensible Markup Language.

(a) eXtensible Modern Language (b) eXtensible Markup List

(c) eXtensible Markov Language (d) None of these

Q. 8 Web browser is a software program that interprets and displays the contents of -------- web pages.

(a) XML (b) HTML (c) static (d) dynamic

Q. 9 Affiliation networks are ----- mode networks that allow one to study the dual perspectives of the actors and the events.

(a) two (b) three (c) four (d) eight

Q. 10 ----- is a tie between two actors and consists of a pair of actors and the tie(s) between them

(a) triad (b) relational tie (c) Dyad (d) graph

Q. 11 --------- is the ratio of the number of relevant records retrieved to the total number of relevant records in the database.

(a) Precision (b) recall (c) link (d) none of these

➤ Answers of Fill in the Blanks

1.	implied	8.	knowledge technology
2.	microblogging	9.	ontologies
3.	machine-readable	10.	knowledge
4.	semantic	11.	pandora
5.	two-mode	12.	Recall
6.	Affiliation	13.	semantic
7.	Hypertext Markup Language	14.	Density

➤ Answers of Multiple Choice Questions

1.	b	2.	b	3.	a	4.	d	5.	a	6.	b
7.	a	8.	b	9.	a	10.	c	11.	b		

❑❑❑

Notes

2 Modelling, Aggregating and Knowledge Representation

Scope of the Syllabus

Ontology and their role in the Semantic Web: Ontology-based knowledge Representation - Ontology languages for the Semantic Web: Resource Description Framework - Web Ontology Language - Modelling and aggregating social network data: State-of-the-art in network data representation - Ontological representation of social individuals - Ontological representation of social relationships - Aggregating and reasoning with social network data - Advanced representations.

2.1 Introduction Knowledge Representation on the Semantic Web

- Semantic Web is used to represent information on and about the current Web using formal languages that computers can process. A major feature of the Semantic Web is the ability to provide definitions for objects and types of objects that are accessable and manipulable from within the Semantic Web.
- A knowledge-based system maintains a knowledge base which stores the symbols of the computational model in form of statements about the domain, and it performs reasoning by manipulating these symbols.
- Knowledge is the information about a domain that can be used to solve problems in that domain. To solve many problems requires much knowledge, and this knowledge must be represented in the computer. As part of designing a program to solve problems, we must define how the knowledge will be represented.
- A representation scheme is the form of the knowledge that is used in an agent. A representation of some piece of knowledge is the internal representation of the knowledge. A representation scheme specifies the form of the knowledge. A knowledge base is the representation of all of the knowledge that is stored by an agent.
- Knowledge representation is the method used to encode knowledge in an intelligent system's knowledge base. The object of knowledge representation is to express knowledge in computer-tractable form, such that it can be used to help intelligent system perform well. Fig. 2.1.1 shows knowledge base and domain.

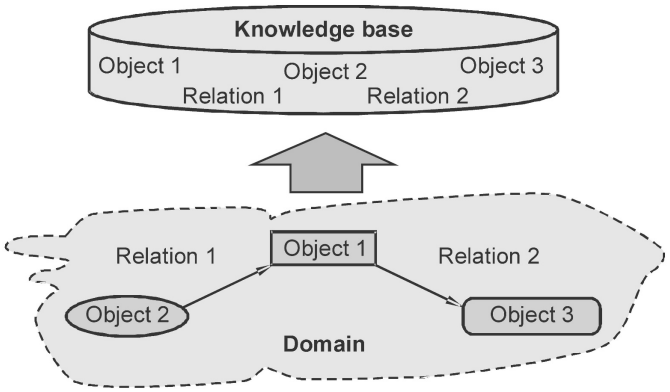


Fig. 2.1.1 : Knowledge base and domain

- A representation consist of four fundamental parts :
  1. A lexical part that determines which symbols are allowed in the representation's vocabulary.
  2. A structural part that describes constraints on how the symbols can be arranged.
  3. A procedural part that specifies access procedures that enable to create descriptions, to modify them, and to answer questions using them.
  4. A semantic part that establishes a way of associating meaning with the description.
- If we look at current Semantic Web technologies and use cases, knowledge representation appears in different forms, the most prevalent of which are based on semantic networks, rules and logic.
- Semantic network structures can be found in RDF graph representations or Topic Maps, whereas a formalization of business knowledge often comes in form of rules with some "if-then" reading, e.g. in business rules or logic programming formalisms. Logic is used to realize a precise semantic interpretation for both of the other forms.
- A semantic network is a graph whose nodes represent concepts and whose arcs represent relations between these concepts. They provide a structural representation of statements about a domain of interest. In the business trips domain, typical concepts would be "Company", "Employee" or "Flight", while typical relations would be "books", "is EmployedAt" or "participatesln".
- In terms of the knowledge representation and reasoning semantic web lets us :
  1. Represent the knowledge
  2. Support search queries on knowledge and matches
  3. Support inference

2.2 Ontology and Their Role in the Semantic Web

- The term ontology is originated from philosophy. In that context it is used as the name of a subfield of philosophy, namely, the study of the nature of existence.

- For the Semantic Web purpose : "An ontology is an explicit and formal specification of a conceptualisation".
- In general, an ontology describes formally a domain of discourse. An ontology consists of a finite list of terms and the relationships between the terms. The terms denote important concepts classes of objects of the domain.
- For example, in a university setting, staff members, students, courses, modules, lecture theatres, and schools are some important concepts.
- In the context of the Web, ontologies provide a shared understanding of a domain. Such a shared understanding is necessary to overcome the difference in terminology.
- Ontologies are useful for improving accuracy of Web searches. Web searches can exploit generalization/ specialization information.
- Ontologies are the core concept of Knowledge Representation in the Semantic Web. RDFS and OWL are languages that can produce such models. The Semantic Web is build on the eXtensible Markup Language (XML), the Resource Description Framework (RDF).

➤ Ontology Languages for the Web

- RDF is a universal language that enables users to describe their own vocabularies. But, RDF does not make assumption about any particular domain. It is up to user to define this in RDF schema.
- RDF is a directed, labeled graph data format for representing information in the Web. Most forms of the query languages contain a set of triple patterns. Triple patterns are like RDF triples except that each of the subject, predicate and object may be a variable.
- RDF Schema is a vocabulary description language for describing properties and classes of RDF resources, with a semantics for generalization hierarchies of such properties and classes.
- OWL is a richer vocabulary description language for describing properties and classes.
- Ontologies are domain models with two special characteristics, which lead to the notion of shared meaning or semantics :
  1. Ontologies are expressed in formal languages with a well-defined semantics.
  2. Ontologies build upon a shared understanding within a community.
- Semantic networks provide a means to abstract from natural language, representing the knowledge that is captured in text in a form more suitable for computation.
- Semantic networks are closely related to another form of knowledge representation called frame systems. In fact, frame systems and semantic networks can be identical in their expressiveness but use different representation metaphors.
- While the semantic network metaphor is that of a graph with concept nodes linked by relation arcs, the frame metaphor draws concepts as boxes, i.e. frames, and relations as slots inside frames that can be filled by other frames. Thus, in the frame metaphor the graph turns into nested boxes.

- The semantic network form of knowledge representation is especially suitable for capturing the taxonomic structure of categories for domain objects and for expressing general statements about the domain of interest.
- Ontologies emerged as an alternative to represent knowledge. However, they have been used to support a great variety of tasks.
- 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.

➤ Why Develop Ontology ?

- Other reasons are as follows :
  1. To enable a machine to use the knowledge in some application.
  2. To enable multiple machines to share their knowledge.
  3. To help yourself understand some area of knowledge better.
  4. To help other people understand some area of knowledge.
  5. To help people reach a consensus in their understanding of some area of knowledge.
- The majority of existing ontologies are 'simple' taxonomies or classifications, i.e., hierarchically organized categories used to classify resources.
- Ontologies with arbitrary relations do exist, but no intuitive and efficient reasoning techniques support such ontologies in general. So we need 'lightweight' ontologies.
- **Lightweight ontology** is typically applied to ontologies that make a distinction between classes, instances and properties, but contain minimal descriptions of them.
- **Heavyweight** ontologies allow to describe more precisely how classes are composed of other classes, and provide a richer set of constructs to constrain how properties can be applied to classes.

2.3 Ontology Languages for the Semantic Web

2.3.1 Resource Description Framework (RDF) and RDF Schema

- The Resource Description Framework (RDF) is a W3C Recommendation since February 2004 which jointly replaced RDF Model and Syntax and RDF Schema.
- RDF provides a means for adding semantics to a document without making any assumptions about the structure of the document and it provides pre-defined modeling primitives for expressing semantics of data.
- RDF is domain-independent and can be used to model both real world objects and information resources. RDF itself is a very primitive modeling language, but it is the basis of more complex languages such as OWL.
- RDF was developed with the motivation to provide web meta data and open information models, to get new information by combining data from several applications and to enable automated processing of web information by software agents.

- RDF is the foundation layer of the Schematic Web. The semantics are enabled in sets of triples, where each triple consists of a subject, a predicate or property and an object, similar to what we have in natural language sentences.
- An RDF statement has three components : **a subject, a predicate and an object**.
  1. The subject is the source of the edge and must be a resource.
  2. The object of a statement is the target of the edge. Like the subject, it can be a resource identified by a URI, but it can alternatively be a literal value like a string or a number.
  3. The predicate of a statement determines what kind of relationship holds between the subject and the object. It too is identified by a URI.
- Fig. 2.3.1 shows an example RDF graph.

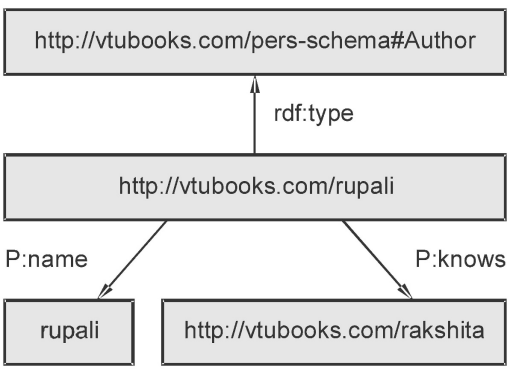


Fig. 2.3.1 : Example RDF graph

- The RDF Vocabulary Description language RDF Schema (RDFS) is an extension to RDF which facilitates the formulation of vocabularies for RDF meta data. While RDF is used to relate resources by means of properties, RDFS introduces the nations of resources classes and their hierarchies.
- The combined use of both RDF and RDFS is often referred to as RDF(S) and provides a simple ontology language for conceptual modeling with some basic interfacing capabilities.
- Fig. 2.3.2 shows an RDF graph, which is a set of RDF triples. A node may be a URI reference or blank, which means that it is a unique node with no separate form of identification. The object node can also be a literal. The property is also a URI.
- An RDF graph is defined as a set of RDF triples. A subgraph of an RDF graph is a subset of the triples in the graph.
- A triple is identified with the singleton set containing it , so that each triple in a graph is considered to be a subgraph.
- A proper subgraph is a proper subset of the triples in the graph. A ground RDF graph is one with on blank nodes.
- The Resource Description Language Schema (RDFS) is a vocabulary language that provides the users to define terms they intend to use in their RDF document, similar to the design in Object Oriented Programming (OOP) languages.

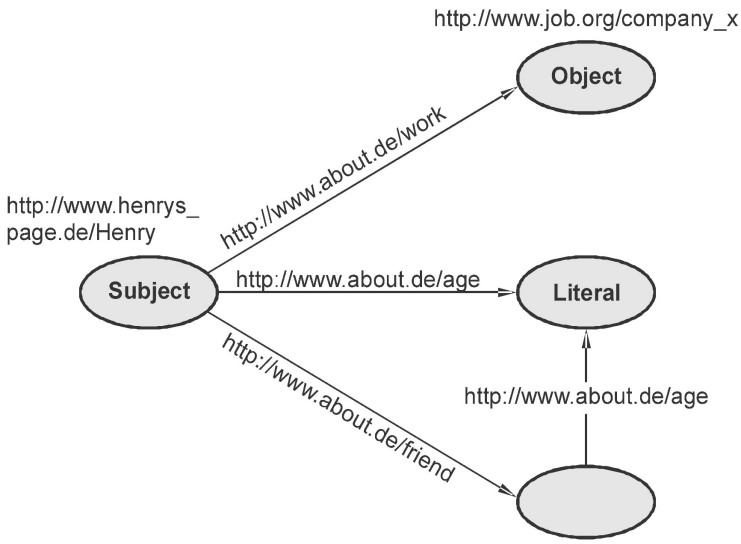


Fig. 2.3.2 : RDF triple relations between resources, literals and blank-nodes

- A set of names is referred to as a vocabulary. The vocabulary of a graph is the set of names which occur as the subject, predicate or object of any triple in the graph.
- The four important RDFS vocabulary definitions are **rdfs:Class, rdf:Property, rdfs:domain and rdfs:range.** They define which nodes are connected through a certain property.
- Things described by RDF expressions are called resources, and are considered to be instances of the class rdfs:Resource. The RDF class rdfs:Resource represents the set called 'Resources' in the formal model.
  1. **rdf:Property :** This represents the subset of RDF resources that are properties i.e. all the elements of the set called 'Properties' the formal model. Example : ex:author rdf:type rdf:Property.
- Then, property ex:author can be used as a **predicate in an RDF triple** such as the following :

  ex:Rupali ex:sportman ex:cricket

- In RDFS property definitions are independent of class definitions. In other words, a property definition can be made without any reference to a class.
- Optionally, properties can be declared to apply to certain instances of classes by defining their domain and range.
  2. **rdfs:Class :** This corresponds to the generic concept of a Type or Category, similar to the notion of a Class in object-oriented programming languages such as Java. When a schema defines a new class, the resource representing that class must have an rdf:type property whose value is the resource rdfs:Class. RDF classes can be defined to represent almost anything, such as Web pages, people, document types, databases or abstract concepts.

- Every RDF model which uses the schema mechanism also (implicitly) includes the core properties. These are instances of the rdf:Property class and provide a mechanism for expressing relationships between classes and their instances or superdasses.

  3. **rdf:type :** This indicates that a resource is a member of a class, and thus has all the characteristics that are to be expected of a member of that class. The value of an rdf:type property for some resource is another resource which must be an instance of rdfs:Class. The resource known as rdfs:Class is itself a resource of rdf:type rdfs:Class.
  4. **rdfs:range :** An instance of ConstraintProperty that is used to indicate the class(es) that the values of a property must be members of. The value of a range property is always a Class. Range constraints are only applied to properties. A property can have at most one range property. It is possible for it to have no range, in which case the class of the property value is unconstrained.
  5. **rdfs:domain :** This is an instance of ConstraintProperty that is used to indicate the class(es) on whose members a property can be used. If a property has no domain property, it may be used with any resource. If it has exactly one domain property, it may only be used on instances of that class. If it has more than one domain property, the constrained property can be used with instances of any of those classes.

- For a property, we can have zero, one, or more than one domain or range statements. **No domain or no range statement :** If no range statement has been made for property P, then **nothing has been said about the values** of this property. Similarly for no domain statement.

➤ Some Utility Properties

1. rdfs:label
2. rdfs:comment
3. rdfs:seeAlso
4. rdfs:isDefinedBy

- **rdfs:label** is an instance of rdf:Property that may be used to provide a human-readable version of a resource's name. The rdfs:domain of rdfs:label is rdfs:Resource. The rdfs:range of rdfs:label is rdfs:Literal. Multilingual labels are supported using the language tagging facility of RDF literals.
- **rdfs:comment** is an instance of rdf:Property that may be used to provide a **human-readable description of a resource**. The rdfs:domain of rdfs:comment is rdfs:Resource. The rdfs:range of rdfs:comment is rdfs:Literal. Multilingual documentation is supported through use of the **language tagging** facility of RDF literals.
- **rdfs:seeAlso** is an instance of rdf:Property that is used to indicate a resource that might provide additional information about the subject resource. A triple of the form S rdfs:seeAlso 0 states that the resource 0 may provide additional information about S. It may be possible to retrieve representations of 0 from the Web, but this is not required. When such representations may be retrieved, no constraints are placed on the format of those representations.

- The rdfs:domain of rdfs:seeAlso is rdfs:Resource. The rdfs:range of rdfs:seeAlso is rdfs:Resource.
- **rdfs:isDefinedBy** is an instance of rdf:Property that is used to indicate a resource defining the subject resource. This property may be used to indicate an RDF vocabulary in which a resource is described.
- RDF Schema descriptions are not prescriptive in the way programming language type declarations typically are.
- RDF Schema provides schema information as additional descriptions of resources, but does not prescribe how these descriptions should be used by an application.
- RDF is a graph data model. RDF data are directed, labeled graphs. A single edge in an RDF graph is a 3-tuple that is called either a statement or triple.
- Triples are organized into named graphs, forming 4-tuples, or quads. RDF resources (nodes), predicates (edges), and named graphs are labeled by URIs.
- The RDFS language is used by various web-based applications for describing meta data, and a number of tools are available that support visual editing and programmatic handling of RDFS descriptions.
- RDF is a language for expressing the information that needs to be processed by applications, so that it can be exchanged without loss of meaning. The data does not need to be stored in RDF but can be created on the fly from relational databases or other non-RDF sources.

➤ **RDF vocabulary :**

Basic constructs	1. rdf:type
	2. rdf:Property
	3. rdf:XMLLiteral
Collections	1. rdf:List
	2. rdf:Seq
	3. rdf:Bag
	4. rdf:Alt
	5. rdf:first
	6. rdf:rest
	7. rdf:nil
	8. rdf:_n
	9. rdf:value
Reification	1. rdf:Statement
	2. rdf:subject
	3. rdf:predicate
	4. rdf:object

➤ RDF Schema Vocabulary

Basic constructs	1. rdfs:domain
	2. rdfs:range
	3. rdfs:Resource
	4. rdfs:Literal
	5. rdfs:Datatype
	6. rdfs:Class
	7. rdfs:subClassOf
	8. rdfs:subPropertyOf
Collections	1. rdfs:member
	2. rdfs:Container
	3. rdfs:ContainerMembershipProperty
Documentation and Reference	1. rdfs:comment
	2. rdfs:seeAlso
	3. rdfs:isDefinedBy
	4. rdfs:label

2.3.2 Web Ontology Language (OWL)

- The Web Ontology Language (OWL) is a knowledge representation language designed by the W3C Web Ontology Working Group for use by applications that need to process web content.
- The OWL standard defines different syntaxes based on RDF(S), XML and proprietary text format. The OWL RDF/XML syntax allows for an encoding of an OWL ontology within the RDF(S) framework in RDF/XML serialisation. The OWL XML presentation syntax pro-vides a more compact XML format for OWL ontologies, independent from RDF(S).
- OWL can also define two types of properties : **object properties and datatype properties**.
- Object properties specify relationships between pairs of resources. Datatype properties, on the other hand, specify a relation between a resource and a data type value; they are equivalent to the notion of attributes in some formalisms.
- OWL Lite, OWL DL and OWL Full are the three levels of OWL.
- OWL Lite supports simple constraints and classification hierarchies which are used to construct taxonomies and thesauri. The cardinality restriction is limited to 0 and 1.
- OWL DL, where DL is the abbreviation for Description Logic, provides maximum expressiveness while retaining computational completeness and decidability.
- Unlike OWL DL, OWL Full does not guarantee computational completeness. Classes described in OWL Full can be treated simultaneously as instances of collections.

- An OWL document can contain the following declarations :
  1. Header
  2. Classes
  3. Complex classes
  4. Individuals
  5. Properties, property characteristics and property restrictions
  6. Ontology mapping

➤ OWL Full vocabulary

Boolean class combinations	1. owl:unionOf
	2. owl:complementOf
	3. owl:intersectionOf
(In)equality of classes and instances	1. owl:equivalentClass
	2. owl:disjointWith
	3. owl:equivalentProperty
	4. owl:sameAs
	5. owl:differentFrom
	6. owl:AllDifferent
	7. owl:distinctMembers
Enumerated types	1. owl:oneOf
	2. owl:DataRange
Property characteristics	1. owl:ObjectProperty
	2. owl:DatatypeProperty
	3. owl:inverseOf
	4. owl:TransitiveProperty
	5. owl:SymmetricProperty
	6. owl:functionalProperty
	7. owl:InverseFunctionalProperty
Property restrictions	1. owl:Restriction
	2. owl:onProperty
	3. owl:allValuesForm
	4. owl:someValuesFrom
	5. owl:hasValue
	6. owl:minCardinality
	7. owl:maxCardinality
	8. owl:cardinality

Ontology versioning	1. owl:versionInfo
	2. owl:priorVersion
	3. owl:backwardCompatibleWith
	4. owl:incompatibleWith
	5. owl:DeprecatedClass
	6. owl:DeprecatedProperty
Ontology metadata	1. owl:ontology
	2. owl:imports
	3. owl:AnnotationProperty
	4. owl:OntologyProperty

2.4 Modelling and Aggregating Social Network Data

- The ability to understand predict human behavior and decision making is an age old problem. Fundamentally, every aspect of our existence, access to resources, and ability to exceed or fail in our endeavors are predicated on interaction with those who, directly or indirectly, make up our environment.
- To a greater or lesser degree all people have the ability to influence aspects of their environment and others within that environment.
- Measures of the strength of connectivity between individuals are termed social closeness where a greater social closeness indicates a stronger influence in the relationship between the individuals. Social closeness is represented as a weight on the edges in a social network graph.
- Maintaining the semantics of social network data is crucial for aggregating social network information, especially in heterogeneous environments where the individual sources of data are under diverse control.
- Semantical representations can facilitate the exchange and reuse of case study data in the academic field of Social Network Analysis.
- After collecting all the information, the first step is to determine the user model dimensions that user model has to cover.
  1. Personal Characteristics (or Demographics) range from basic information like gender or age to more social ones like relationship status.
  2. Interests and Preferences in an adaptive system usually describe the users interest in certain items. Items can be e.g. products, news or documents.
  3. Needs and Goals : When using computer systems, users usually have a goal they want to achieve. Such goals can be to satisfy an information need or to buy a product.

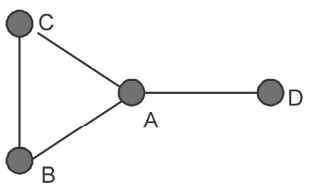
4. Knowledge and Background describe the users knowledge about a topic or system. It is used in educational systems to adapt the learning material to the knowledge of a student, display personalized help texts or tailor descriptions to the technical background of a user.

2.4.1 State-of-the-art in Network Data Representation

- For modelling the social network data, graph is used. In this representation, where the nodes represents individuals and the edges represent binary social relationships. Social network studies build on attributes of nodes and edges.
- A common technique that analysis use to draft a sociogram is to construct it around the circumference of a circle. The circle helps organize the data, but the order in which analysts place the points is determined only by their attempt to keep the number of lines connecting the various points to a minimum.
- Typically, researchers using this technique engage in a trial-and-error drafting process until they reach an aesthetically pleasing result.
- While such a process can make the structure of relations clearer, the relations between the sociogram's points reflect no specific mathematical properties. The points are arranged arbitrarily and the distances between them are meaningless.
- Pajek, which is Slovenian for "Spider," is a network analysis and graph drawing program that has specifically been designed to handle extremely large data sets. These are text-based formats which have been designed in a way so that they can be easily edited using simple text editors.
- Pajek has six data structures to implement the algorithms : network, permutation, vector, cluster, partition and hierarchy and it is based on transformations that support transitions between this data.

➤ Pajek's Data/NET

- The basic data format is the .net file. Let us consider following example with output as graph.

*Vertices 4 1 "A" 2 "B" 3 "C" 4 "D" *Edges 1 2 1 1 3 1 2 3 1	The file defining the graph 

- The network described on the file.net is shown in Fig. 2.4.1.

*Vertices n	n is the number of vertices
1 "$name_1$"	the label of vertex 1 is $name_1$
. . .	
n "$name_n$"	
*arcs	
i j v_{ij}	the arc from i to j has value v_{ij}
. . .	
*edges	
p q v_{pq}	the edge from p to q has value v_{pq}
. . .	

Fig. 2.4.1 : File.net structure

- All three types of files have the same structure.

*Vertices n	n is the number of vertices
v_1	vertex 1 has value v_1
. . .	
v_n	

- The input files can be created / modified by any text editor. Leave no blank lines. Next to the .net files, there can be
  1. .clu files with nominal data (partitions),
  2. .vec file with numeric data,
  3. .per files with permutations (orderings).
- All have the same structure. It is shown in Fig. 2.4.2.

Partition

```
*Vertices 4
1
1
1
2
```

Vector

```
*Vertices 4
0.5
1
1.5
40
```

Permutation

```
*Vertices 4
4
2
1
3
```

Fig. 2.4.2 : File structure

- **Graph drawing :** To draw a graph , choose the network / partition / vector that you wish to work with, and use the Draw menu. If you wish to use multiple colors for the vertices, they must be defined in a partition, and the command to be used is Draw-Partition (Ctrl-P).
- All types of data can be combined into a single file: -Pajek's project file.paj.
- Pajek supports also two-mode and temporal networks.

➤ UCINET :

- It is a windows software package used for the analysis of social network data, has been developed by Steve Borgatti, Martin Everett and Lin Freeman.
- This package provides the tools to analyze 1-mode or 2-mode data, and it can handle a maximum of two millions nodes, but in practice most of the procedures are too slow to run networks larger than 5000 nodes.
- UCINET contains dozen of network analytical tools, such as : centrality measure, subgroup identification, role analysis, elementary graph theory and permutation-based statistical analysis.
- The software can perform a wide variety of data transformation such as : sub-graphs and sub-matrices, merging datasets, permutation and sorts, transposing and reshaping, recodes, linear transformations, symmetrizing, geodesic distances and reachability, aggregation, normalizing and standardizing, mode changes.

➤ GraphML :

- Graph drawing tools, like all other tools dealing with relational data, need to store and exchange graphs and associated data. Despite several earlier attempts to define a standard, no agreed-upon format is widely accepted and, indeed, many tools support only a limited number of custom formats which are typically restricted in their expressibility and specific to an area of application.
- GraphML format represents an advancement over the previously mentioned formats in terms of both interoperability and extensibility.
- GraphML originates from the information visualization community where a shared format greatly increases the usability of new visualization methods.
- GraphML is therefore based on XML with a schema defined in XML Schema. This has the ad-vantage that GraphML files can be edited, stored, queried, transformed etc. using generic XML tools.

```
<graphm1>
    <graph edgedefault="directed">
        <node id="v1"/>
        <node id="v2"/>
        <node id="v3"/>
        <node id="v4"/>
        <edge source="v1" target="v2"/>
        <edge source="v1" target="v3"/>
        <edge source="v2" target="v4"/>
        <edge source="v2" target="v4"/>
        directed="false"/>
    </graph>
</graphm1>
```

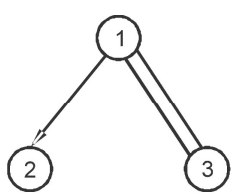


Fig. 2.4.3 : Graph and its representation in GraphML

- Besides GraphML there is a multitude of file formats for serializing graphs. Among the simplest ones are direct ASCII-based codings of tables (matrices) or lists, such as tab-separated value files. Specific instances of these include UCINET's *.dl files and Pajek's *.net .
- Fig. 2.4.3 shows graph and its representation in GraphML.

2.5 Ontological Representation of Social Individuals

- Knowledge management ensures the development and application of all types of relevant knowledge in a company in order to improve their ability to solve problems and contribute to the sustainability of their competitive advantages.
- It is necessary to identify those who are a source of valuable information and support the transformation of organizational knowledge to some structured form that can be processed.
- The purpose of the ontology network proposed is to model semantic information concerning the knowledge objects.
- Framework for modelling and characterizing social relationships for two main reasons :
  1. To support the automated integration of social information on a semantical basis and
  2. To capture established concepts in Social Network Analysis.
- Due to the low-cost of friendship identification in online social networks and the variance of link information in electronic communication networks, the resulting networks contain both strong and weak ties, with little or no information to differentiate between the two ends of the spectrum.
- Since pairs of individuals with strong ties (e.g., close friends) are likely to exhibit greater similarity than those with weak ties , treating all relationships equally will increase the level of noise in the learned models.
- Personal content applications use social ontology for the annotation and retrieval of multimedia documents.
- Tidepool from Immuexa is a personal application that lets users collect and organize their 'personal memories' in digital form.
- Tidepool adds RDF-based annotations (FOAF, in specific) to multimedia documents and allows users to navigate through content by browsing through ontological terms.
- Tidepool also combines this technology with instant messaging for annotating images collectively and sharing content and metadata with friends and family.
- The main idea is that the representation of social relationships needs to be fine-grained enough so that it is possible to capture all the detail from the individual sources of information.
- Network analysis contains rich set of vocabulary for characterizing social relationships. Following list contains some of them.

Vocabulary	Description
Sign	• A relationship can represent both positive (i.e. like) and negative (i.e. hate) attitudes. • This type of relationship helps for study of balance within social networks.
Strength	• The notion of tie strength was first introduced by Granovetter. • It is complex construct of several characteristics of social relations.
Provenance	• A social relationship may be seen differently by the individual participants of the relationship. • Only one member shows the relationship.
Relationship history	• Social relationships come into existence by some event. • Social relationships has lifecycle and it start because of some event and ends when relationship change.
Relationship roles	• Social relationship may have a number of social roles associated with it.

- Online social networks (OSNs) often consist of more than just a record of social network ties. Typically online communities contain ancillary interaction information among the users that can be used to improve modeling.
- Treating all relationships as equal will increase the level of noise in a learned model and degrade performance.
- Recently, social interactions through web 2.0 social platforms have raised lots of attention in the semantic web community. Several ontologies are used to represent social networks. FOAF is used for describing people profiles, their relationships and their activities online.
- The properties of the RELATIONSHIP ontology specialize the "knows" property of FOAF to type relationships in a social network more precisely.
- All these ontologies can be used and extended to link and reuse scenarios and data from web 2.0 community sites. RDF based descriptions of social data provide a rich typed graph and offer a much more powerful and significant way to represent online social networks than traditional models of SNA.
- Social relations could be represented as n-ary predicate reification.

➤ Direct Binary Relations

- A common way to represent n-ary facts is to simply decompose them directly into binary relations between two participants. But here important information may be lost.
- For instance, given a triple with property wasMarriedOnDate and two triples with gotMarriedTo, we cannot be sure to which marriage the given time span applies.
- **Example : Suresh MarriedOnDate 1998**

  RDF Reification

- The RDF proposes RDF reification, which introduces a new identifier for a statement and then describes the original RDF statement using three new triples with **subject, predicate, and object properties**.
- RDF reification is used to attach additional information to the event represented by the original RDF triple.
- RDF Reification example :

Suresh	marries	Deepa
s	type	statement
s	subject	Suresh
s	property	marries
s	object	Deepa
s	time	1998

- In RDF reification, an entity is defined that stands for a whole triple so that additional triples can be used to describe the reified triple as a unit that represents a statement.
- However, in the context of event semantics, reification is used to denote the process by which an entity is defined that refers to the event, process, situation, or more generally, frame, evoked by a property or binary relation.
- Social relations are socially constructed objects : they are constructed in social environments by assigning a label to a common pattern of interaction between individuals.

2.6 Aggregating and Reasoning with Social Network Data

- Semantic technologies have proved evidence of efficient implementations on stream data domains. OWL ontologies have been widely used for modeling stream data domains.
- Querying these knowledge bases has been merely done by SPARQL extensions.
- Managing the knowledge bases and reasoning with background and streaming data is merely done by rule systems

➤ Task of aggregation

- It requires capturing the domain-specific knowledge of when to consider two instances to be the same. In practice, only a limited part of the knowledge regarding instance equality can be captured in RDF or OWL.
- Determining equality is often considered as applying a threshold value to a similarity measure, which is a weighted combination of the similarity of certain properties.

➤ Representing identity

- Multiple identifiers can be represented in RDF in two separate ways : new separate resource and use the identifiers as URIs for these resources.
- Once separate resources are introduced for the same object, the equality of these resources can be expressed using the **owl:sameAs** property. The other alternative is to chose one of the identifiers and use it as a URI.

- RDF and OWL allow to identify resources and to represent their (in)equality using the **owl:sameAs** and **owl:differentFrom** properties
- Recall that in OWL, all relationships are binary, so facts are expressed as **subject-predicate-object** sentences. The set of instances from which the subjects are drawn is the domain of the predicate, and the set of instances from which the objects are drawn is the range of the predicate.
- In OWL, binary predicates are called properties. Object properties relate entities to entities, while data properties relate entities to literals.
- If the domain and range are the same or at least compatible, it is meaningful to compare subject and object instances.
- In mathematics, a relation R on a set A is said to be reflexive over its domain if and only if xRx for each element x in A.
- OWL allows object property expressions to be declared globally reflexive by characterizing them as reflexive properties. For example, if we restrict the domain of individuals (owl:Thing) to persons, and we agree that each person knows himself/herself, then the "knows" predicate may be declared to be reflexive.
- In OWL, if a predicate is known to be reflexive, then any instance in its domain may be inferred to bear the relationship to itself.
- While OWL does not directly support local reflexivity in this sense, it does include an ObjectHasSelf restriction to declare a subset of a predicate's domain where each individual in that subset bears the relation to itself.
- For example, suppose we declare the predicate "likes" between people. Some but not all people like themselves, so "likes" is not reflexive. However, we may define SelfLiker as a subclass of Person, where each person in SelfLiker does like himself/herself.

➤ Forward versus backward chaining

- Rules can be used either in a forward-chaining or backward-chaining.
- Forward chaining means that all consequences of the rules are computed to obtain what is called a complete materialization of the knowledge base.
- Advantages : Queries are fast to execute
- Disadvantage : Required extra space
- Backward-chaining, rules are executed "backwards" and on demand, i.e. when a query needs to be answered.
- The drawback of backward-chaining is longer query execution times.

➤ RDF data smushing

- Smushing is the process of normalising an RDF dataset in order to unify a priori different RDF resources which actually represent the same thing.
- The application which executes a data smushing process is called a smusher.

- The process comprises two stages :
  1. Redundant resources are identified;
  2. The dataset is updated to reflect the recently acquired knowledge.
- The latter is usually achieved by adding new triples to the model to relate the pairs of redundant resources. The owl:sameAs is often used for this purpose, although other properties without built-in logic interpretations can be used as well.

2.6.1 Advanced Representations

- Temporal logic is required to formalize ontology versioning, which is also called change management.
- Extending logic with probabilities is also a natural step in representing our problem more accurately.
- RIF will most likely include first-order logic (FOL).
- Extending logic with probabilities is also a natural step in representing problem more accurately.
- When combining similarities, use weights to express the relevance of certain properties. Such weights can be determined adhoc or can be learnt from training data using machine learning. The resulting probability could be represented in a probabilistic logic.

2.7 Questions with Answers

2.7.1 Two Marks Questions with Answers

Q. 1 What is Knowledge representation ?

Ans. : Knowledge representation is a branch of Artificial Intelligence that provides access to a structured collection of information and a set of inference rules. The information and rules can then be used for automated reasoning, e.g. with the help of software agents.

Q. 2 Define ontology.

Ans. : Ontology is a hierarchically structured set of terms to describe a domain that can be used as a skeletal foundation for a knowledge base.

Q. 3 What is lightweight ontology ?

Ans. : Lightweight ontology is applied to ontologies that make a distinction between classes, instances and properties, but contain minimal descriptions of them

Q. 4 What is heavyweight ontology ?

Ans. : Heavyweight ontologies allow to describe more precisely how classes are composed of other classes, and provide a richer set of constructs to constrain how properties can be applied to classes.

Q. 5 Explain RDF reification.

Ans. : In RDF reification, an entity is defined that stands for a whole triple so that additional triples can be used to describe the reified triple as a unit that represents a statement

Q. 6 What is forward chaining ?

Ans. : Forward chaining means that all consequences of the rules are computed to obtain what is called a complete materialization of the knowledge base.

Q. 7 What is backward chaining ?

Ans. : With backward-chaining, rules are executed "backwards" and on demand, i.e. when a query needs to be answered.

Q. 8 What is knowledge-based system ?

Ans. : A knowledge-based system maintains a knowledge base which stores the symbols of the computational model in form of statements about the domain, and it performs reasoning by manipulating these symbols.

Q. 9 Which are three components of RDF statement ?

Ans. : An RDF statement has three components : **a subject, a predicate, and an object**.

Q. 10 What is smushing ?

Ans. : Smushing is the process of normalising an RDF dataset in order to unify a priori different RDF resources which actually represent the same thing.

Q. 11 Define OWL.

Ans. : The Web Ontology Language (OWL) is a knowledge representation language designed by the W3C Web Ontology Working Group for use by applications that need to process web content

Q. 12 What are the three levels of OWL?

Ans. : OWL Lite, OWL DL and OWL Full are the three levels of OWL.

Q. 13 Define OWL Lite and OWL DL.

Ans. :

- OWL Lite supports simple constraints and classification hierarchies which are used to construct taxonomies and thesauri. The cardinality restriction is limited to 0 and 1.
- OWL DL, where DL is the abbreviation for Description Logic, provides maximum expressiveness while retaining computational completeness and decidability.

Q. 14 What is OWL full?

Ans. : OWL full allows free mixing of OWL with RDF Schema. It uses all the OWL languages primitives. Unlike OWL DL, OWL Full does not guarantee computational completeness. Classes described in OWL Full can be treated simultaneously as instances of collections.

Q. 15 What is FOAF (Friend of a friend)?

Ans. :

- FOAF is used for describing people profiles, their relationships and their activities online. Its aim is to create a linked information system about people, groups, companies and other kinds of thing.
- If people publish information in FOAF document format, machines will be able to make use of that information.
- If those files contain "see also" references to other such documents in the Web, we will have a machine-friendly version of today's hypertext web. FOAF documents are usually represented in RDF.

Q. 16 What is Direct Binary Relations?

Ans. : A common way to represent n-ary facts is to simply decompose them directly into binary relations between two participants. But here important information may be lost. For instance, given a triple with property *wasMarriedOnDate* and two triples with *gotMarriedTo*, we cannot be sure to which marriage the given time span applies

2.7.2 Fill in the Blanks

Q. 1 Ontology refers to the concept of ----------.

Q. 2 SPARQL is a query language developed for the ---------- model.

Q. 3 Ontology applications have been developed for ---------- natural language processing.

Q. 4 Ontology on the web will combine a -------- with a set of inference rules

Q. 5 ---------- can be considered as either a retrieval problem or a clustering problem

Q. 6 OWL is a richer ---------- language for describing properties and classes.

Q. 7 RDF ------- is used to attach additional information to the event represented by the original RDF triple

Q. 8 RDF is the foundation layer of the ----------.

Q. 9 An RDF statement has three components: -------, ------- and an --------.

Q. 10 Ontology is a ---------- structured set of terms to describe a domain that can be used as a skeletal foundation for a knowledge base.

Q. 11 In OWL, binary predicates are called ----------.

Q. 12 Semantic networks are closely related to another form of knowledge representation called --------- system.

2.7.3 Multiple Choice Questions

Q. 1 OWL can also define two types of properties: ---------- and----------

(a) Object , datatype (b) XML, HTML

(c) reification, class (d) None of these

Q. 2 RDFS stands for ----------

(a) Record Description Language Schema

(b) Resource Description Language Schema

(c) Resource Description Layer Schema

(d) Resource Description Language System

Q. 3 --------fromImmuexa is a personal application that lets users collect and organize their 'personal memories' in digital form.

(a) Pajek (b) GraphML (c) Tidepool (d) All of these

Q. 4 ---------- istherefore based on XML with a schema defined in XML Schema.

(a) HTML (b) Pajek (c) GraphML (d) None of these

Q. 5 FOAF profiles can be linked together to form networks of web-based profiles.

(a) FOAF (b) OWL (c) RDF (d) network

Q. 6 FOAF profiles are typically posted on the personal website of the user and linked from the user's homepage with the ---------- META tag.

(a) XML (b) OWL (c) RDFS (d) HTML

Q. 7 Very low reasoning support is provided by ----------

(a) OWL Full (b) OWL DL (c) OWL Partial (d) OWL Lite

➤ Answers of Fill in the Blanks

1.	existence	7.	reification
2.	RDF data	8.	Semantic Web
3.	knowledge management	9.	Subject, predicate, object
4.	taxonomy	10.	hierarchically
5.	Smushing	11.	properties
6.	vocabulary description	12.	frame

➤ Answers of Multiple Choice Questions

1.	a	2.	b	3.	c	4.	c	5.	a	6.	d	7.	a

❑❑❑

3 Extraction and Mining Communities in Web Social Networks

Scope of the Syllabus

Extracting evolution of Web Community from a Series of Web Archive - Detecting communities in social networks - Definition of community - Evaluating communities - Methods for community detection and mining - Applications of community mining algorithms - Tools for detecting communities social network infrastructures and communities - Decentralized online social networks - Multi-Relational characterization of dynamic social network communities.

3.1 Extracting Evolution of Web Community from a Series of Web Archive

- Recent advances in storage technology make it possible to store a series of large Web archives. It is now an exciting challenge for us to observe evolution of the Web.
- A web community is a set of web pages created by individuals or associations with a common interest on a topic.
- The data stream has gained importance in recent years because of the greater ease in data collection methodologies resulting from advances in hardware technology.
- For extracting information, we used four web communities used by Japanese web archives crawled from 1999 to 2002 with 119 million pages in total.
- The web community chart is a graph that consists of web communities as nodes and weighted edges between related communities. The weight of each edge represents the relevance of communities at both ends.
- Most research on web communities is based on the notion of authorities and hubs proposed by Kleinberg. An authority is a page with good contents on a topic, and is pointed to by many good hub pages.
- A hub is a page with a list of hyperlinks to valuable pages on the topic, i.e. points to many good authorities. Fig. 3.1.1 shows hub and authorities.

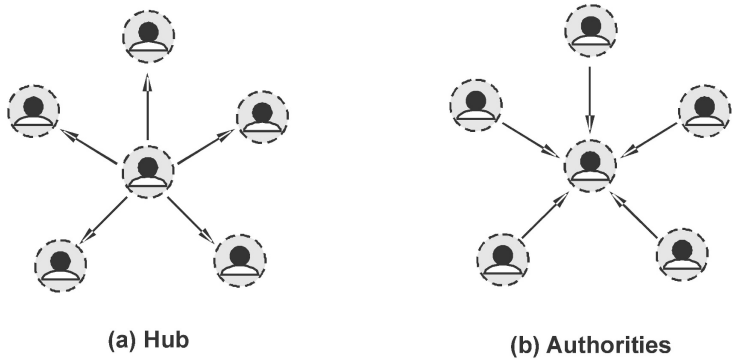
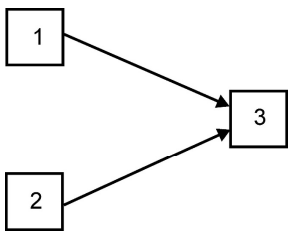


Fig. 3.1.1

- HITS is an algorithm that extracts authorities and hubs from a given sub-graph of the Web with efficient iterative calculation. Evolution metrics are defined to measure the degree of community evolution.
- The main advantage of web community chart is the existence of relevance between communities. Combining evolution metrics and relevance, we can locate evolution around a particular community.
- Tyler et al uses following method for Community Extraction
  1. A graph theoretic algorithm for discovering communities
  2. The graph is broken into connected components and each component is checked to see if it is a community
  3. If a component is not a community then iteratively remove edges with highest betweenness till component splits. Betweenness is recomputed each time an edge is removed
  4. The order of in which edges are removed affects the final community structure
  5. Since ties are broken arbitrarily, this affects the final community structure
  6. In order to ensure stability of results, the entire procedure is repeated i times and the results from each iteration are aggregated to produce the final set of communities
- Girvan and Newman (2002) use a similar algorithm to analyze community structure in social and biological networks

Example 3.1.1 : Let us consider a very simple graph :



The adjacency matrix of the graph is

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

With transpose

$$A^t = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Assume the initial hub weight vector is :

$$u = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Ans. :

We compute the authority weight vector by :

$$v = A^t \cdot u = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

Then, the updated hub weight is :

$$u = A \cdot v = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

This already corresponds to our intuition that node 3 is the most authoritative, since it is the only one with incoming edges, and that nodes 1 and 2 are equally important hubs. If we repeat the process further, we will only obtain scalar multiples of the vectors v and u computed at step 1. So the relative weights of the nodes remain the same.

3.2 Detecting Communities in Social Networks

- Network is used to represent real world entity. For example social network is connected with friendships or co-authors. In most of the example, real social network contains two parts : denser and sparser part.
- In denser sub-network, group of peoples are closely connected to each other. This type of denser sub-network called as communities.
- Detecting communities from given social networks are practically important for the following reasons:
  1. For information recommendation, communities are used. In communities, members have similar preferences and tests.
  2. Communities will help us understand the structures of given social networks.
  3. Communities will play important roles when we visualize large-scale social networks
- Communities, also known as modules and clusters, are sets of nodes which are relatively more connected, and are believed to be the intrinsic structures in networks in the nature.
- Nodes in the same community often share interesting properties such as a common function, interest, or purpose. Thus, community detection is one of the most important problems in network analysis.

➤ Why Community Detection ?

1. Communities in a citation network might represent related papers on a single topic;
2. Communities on the web might represent pages of related topics;
3. Community can be considered as a summary of the whole network thus easy to visualize and understand.
4. Sometimes, community can reveal the properties without releasing the individual privacy information.

3.3 Definition of Community

- Definition is divided into three parts:
  1. Local definitions
  2. Global definitions
  3. Definitions based on vertex similarity.

3.3.1 Local Definition

- The attention is focused on the vertices of the sub-network under investigation and on its immediate neighborhood.
- A local definition of community is divided into two types : **self-referring ones and comparative ones**.
- The examples of self referring definitions are clique , n-clique and k-plex.
  1. Clique : A maximal sub-networks where each vertex is adjacent to all the others.
  2. n-clique : A maximal sub-network such that the distance of each pair of vertices is not larger than n
  3. k-plex : A maximal sub-network such that each vertex is adjacent to all the others except at most k of them.
- The examples of comparative definitions are LS set and weak community
- LS set : sub-network where each vertex has more neighbors inside than outside of the sub-network
- Weak community : The total degree of the vertices inside the community exceeds the number of edges lying between the community and the rest of the network.
- Fig. 3.3.1 shows a graph and a listing of the cliques contained in it. The sub-graphs are in fact cliques, and that there are no remaining cliques in the graph. Notice that cliques in a graph may overlap. The same node or set of nodes might belong to more than one clique.

  Cliques = {1, 2, 3}, {1, 3, 5} and {3, 4, 5, 6}

- For example, in figure node 3 belongs in all three cliques. Also, there may be nodes that do not belong to any cliques (for example node 7). However, no clique can be entirely contained within another clique, because if it were the smaller clique would not be maximal.

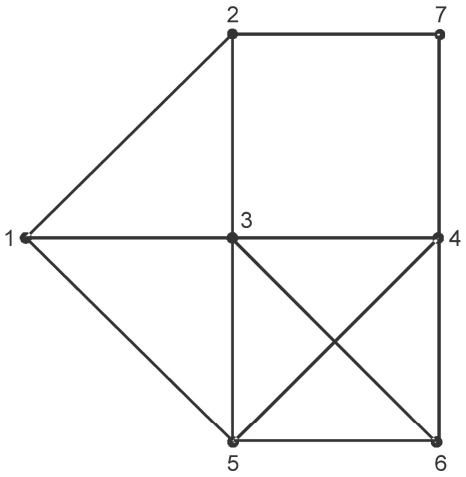


Fig. 3.3.1 : Graph and cliques

3.3.2 Global Definitions

- A global definition of community is related to a sub-network with respect to the network as a whole. It starts from a null model.
- Network which matches the original network in some of its topological features, but which does not display community structure. Then, the linking properties of sub-networks of the initial network are compared with those of the corresponding sub-networks in the null model. If there is a wide difference between them, the sub-networks are regarded as communities.
- Null model is designed by using randomness in the distribution of edges among vertices. The most popular null model is that proposed by Newman and Girvan.
- Null model consists of a randomized version of the original network, where edges are rewired at random, under the constraint that each vertex keeps its degree. This null model is the basic concept behind the definition of modularity, a function which evaluates the goodness of partitions of a network into communities.

3.3.3 Definitions Based on Vertex Similarity

- Last type of definition is based on an assumption that communities are groups of vertices which are similar to each other. To evaluate the similarity between each pair of vertices, some calculation is used.
- Similarity measures are based on hierarchical clustering. Hierarchical clustering is a way to find several layers of communities that are composed of vertices similar to each other.
- Repetitive merges of similar vertices based on some quantitative similarity measures will generate a structure shown in Fig. 3.3.2. This structure is called dendrogram.
- Decompose data objects into a several levels of nested partitioning (tree of clusters), called a **dendrogram**. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

- Highly similar vertices are connected in the lower part of the dendrogram. Subtrees obtained by cutting the dendrogram with horizontal line correspond to communities. Communities of different granularity will be obtained by changing the position of the horizontal line.

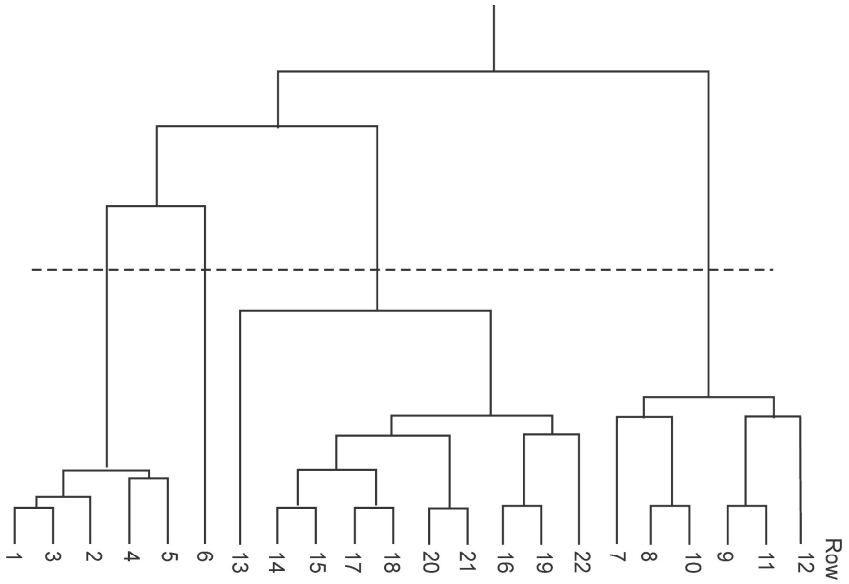


Fig. 3.3.2 : Dendrogram

- The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters.
- Each joining of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters.
- A cross-section of the tree at any level, as indicated by the dotted line, will give the communities at that level

3.4 Evaluating Communities

- Various methods are used for partitioning given network into communities. It is necessary to establish which partition exhibit a real community structure.
- Quality function supports for finding good partitions. The most popular quality function is the modularity.
- Newman and Girvan were among the first to address this issue and proposed modularity to quantify the strength of community structure.
- This metric, based on the intuition that nodes within the same community should be more tightly connected than they would be by chance, has been adopted for a variety of uses including the validation and comparison of community structures, but also as an objective function for optimization algorithms to identify communities.

- Fig. 3.4.1 shows a small network with community structure. In this case there are three communities, denoted by the dashed circles, which have dense internal links but between which there are only a lower density of external links.

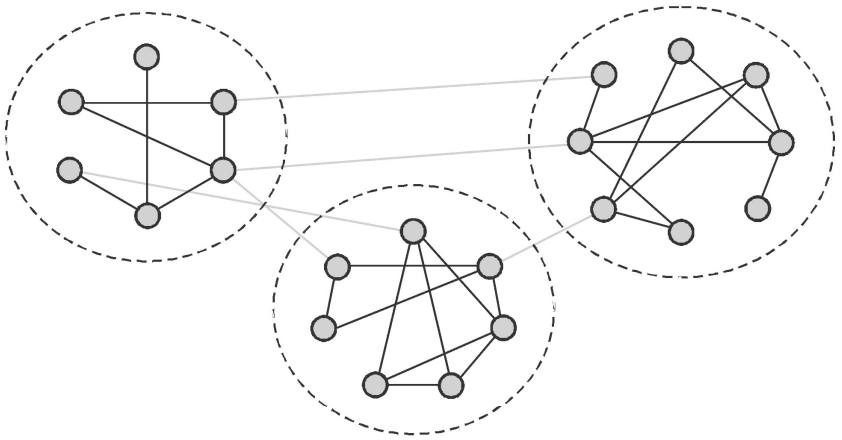


Fig. 3.4.1 : Small network with community structure

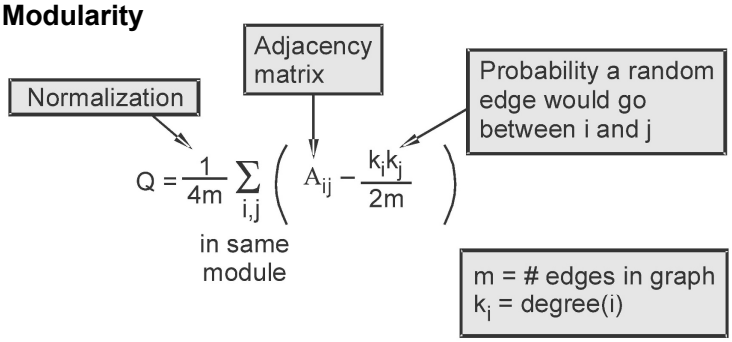
- A graph can be split into communities in numerous ways, i.e. for each graph there are many possible community structures. In the simple case, a community structure is defined as a graph partition into a set of node sets $C = \{C_i\}$.
- To provide a measure of the quality of a community structure, we make use of modularity.
- Modularity quantifies the extent to which a given graph partition into communities presents a systematic tendency to have more intra-community links than the same community structure would present if the links would be rewired under ER (Erdos-Renyi) graph model.
- Modularity (Q) is defined in several ways.

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right)$$

Where e_{ii} = Probability edge is in module i

a_i^2 = Probability a random edge would fall into module i

➤ **Another View of Modularity**



- Modularity measures the strength of a community partition by taking into account the degree distribution. A larger value indicates a good community structure
- One advantage of modularity is that it can be computed using only connectivity of the network, in the absence of any node labels or other information. However, this property can also be considered a weakness because modularity is unable to incorporate metadata (e.g. node labels) even if it is available.
- Modularity measures internal and not external connectivity, but it does so with reference to a randomized null model.
- The modularity can be either positive or negative. Positive values indicate the possible presence of community structure

3.5 Methods for Community Detection and Mining

- The classical methods for dividing given networks into sub-networks are graph partitioning, hierarchical clustering, and k-means clustering.
- All these methods depend upon the numbers of clusters or their size in advance. It is necessary to find suitable methods that have abilities of extracting complete information about the community structure of networks.
- The methods for detecting communities are roughly classified into the following categories:
  1. Divisive algorithms
  2. Modularity optimization
  3. Spectral algorithms

3.5.1 Divisive Algorithm

- Simple method to identify communities in a network is to find the edges that can connect vertices of different communities and remove them, so that the communities get disconnected from each other.
- Newman-Girvan algorithm was has two best features :
  1. They involve iterative removal of edges from the network to split it into communities, the edges removed being identified using "betweenness" measure which represents number of shortest paths between pair of nodes that pass through the links
  2. These measures are recalculated after each removal.
- Newman-Girvan algorithms are highly effective at discovering community structure in both computer-generated and real-world network data, and they can be also used for complex structure of networked systems. Fig. 3.5.1 shows detecting communities based on edge betweenness.
- It uses the idea that "bridges" between communities must have high edge betweenness. The edge with higher betweenness tends to be the bridge between two communities.
- The edge betweenness of an edge is the number of shortest paths between pairs of vertices run along it. Iteratively removing the edges with highest betweenness, we can determine a hierarchical tree and then communities.

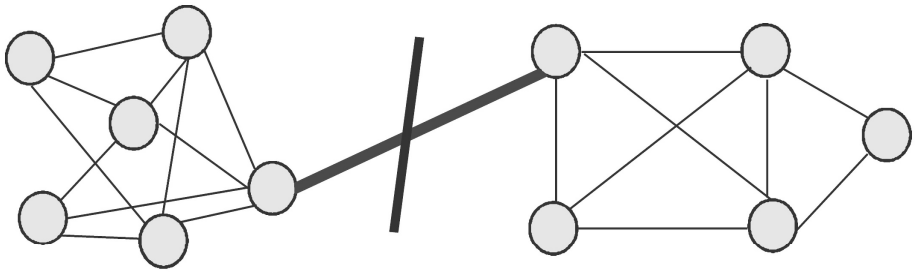


Fig. 3.5.1 : Detecting communities based on edge betweenness

3.5.2 Modularity Optimization

- An exhaustive optimization of modularity is impossible since there are huge numbers of ways to partition a network. It has been proved that modularity optimization is an NP-hard problem.
- There are currently several algorithms that are able to find fairly good approximations of the modularity maximum in a reasonable time. One of the famous algorithms for modularity optimization is CNM algorithm proposed by Clauset et al.
- Another example of the algorithms are greedy algorithms and simulated annealing.
- Simulated annealing was proposed by Kirkpatrick et al. who noted the conceptual similarity between global optimization and finding the ground state of a physical system.

➤ Simulated Annealing

- To get global optimization, simulated annealing is probabilistic procedure used in different fields and problems. This procedure consists of the space of possible states looking for the maximum global optimum of a function F.
- The standard implementation combines two types of moves : local moves, where a single node is shifted from one cluster to another randomly; and global moves, which consist of mergers and splits of communities

3.5.3 Spectral Algorithms

- Spectral algorithms are to cut given network into pieces so that the number of edges to be cut will be minimized. Spectral graph bi-partitioning is example of this category.
- There are two categories of spectral algorithms for maximizing modularity: one is based on the modularity matrix and the other is based on the Laplacian matrix of a network.
- In this new method which is based on spectral clustering, the correctness and conductivity function are used to calculate the value of community detection.
- Spectral methods for community detection rely upon normalized cuts for clustering. A cut partitions a graph into separate parts by removing edges; it shown in Fig. 3.5.2.
- Spectral clustering partitions a graph into two sub-graphs by using the best cut such that within community connections are high and across-community connections are low.

- It can be shown that a relaxation of this discrete optimization problem is equivalent to examining the eigen-vectors of the Laplacian of the graph. For this research, divisive clustering was used, recursively partitioning the graph into communities by "divide and conquer" methods.
- One of the most common methods for community detection is bisection method which is based on the spectral clustering mainly uses Graph Spectral Theory.

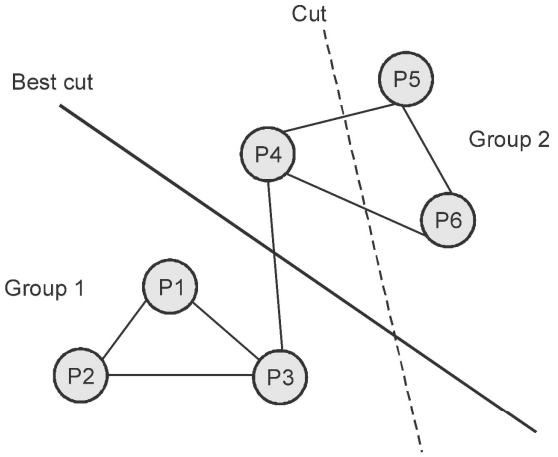


Fig. 3.5.2 : Spectral clustering of a graph relies on recursive binary partitions

- Here the proposed algorithm uses the eigen value distribution of the Laplacian matrix to estimate the number of communities and the k-means algorithm is used for clustering.
- The drawback of this algorithm is that it is applicable to the network graph which can be clearly divided into two communities, but the number of communities which are shared, cannot be detected.
- Applications of community detection are as follows :
  1. Recommendation system
  2. Social network role detection
  3. Functional module in biological networks
  4. Graph coarsening and summarization
  5. Network hierarchy inference

3.6 Applications of Community Mining Algorithm

- Following are the community mining algorithms :
  1. Network reduction
  2. Discovering scientific collaboration groups from small networks
  3. Mining communities from dynamic and distributed networks

➢ 1. Network reduction

- Analysing social networks, network reduction method is used. For example, one author may write many paper for online sits.
- Suppose for bibliography, 300 author can write 415 papers. Here it is possible to reduce co-author network. Clustered co-author network can be reduced into a much smaller one by condensing each community as one node.
- Dendrogram corresponding to the original co-author network can be built.

➢ 2. Discovering scientific collaboration groups from social networks

- Social network like Flink describe the scientific collaborations of more than 600 semantic web researchers. For considering Social network analysis viewpoints, some questions are as follows:
  a) Which researcher would more likely to collaborate with each other.
  b) What are the main reasons that bind them together?
- By applying community mining techniques, we can solve above questions.

➢ 3. Mining communities from dynamic and distributed networks

- Application involve distributed and dynamically evolving networks. Resources and controls are not only decentralized but also updated frequently.
- We need to find hidden communities from distributed and dynamic networks.

3.7 Tools for Detecting Communities Social Network Infrastructures and Communities

- Wide varieties of tools are available for detecting communities. These tools are divided into two categories :
  1. detecting communities from large-scale networks
  2. interactively analyzing communities from small network

3.7.1 Tools for Large-Scale Networks

- The ability to find communities within large networks in some automated fashion could be of considerable use. Clauset et al. propose CNM algorithm of community detection based on modularity optimization.
- Communities in a web graph for instance might correspond to sets of web sites dealing with related topics , while communities in a biochemical network or an electronic circuit might correspond to functional units of some kind.

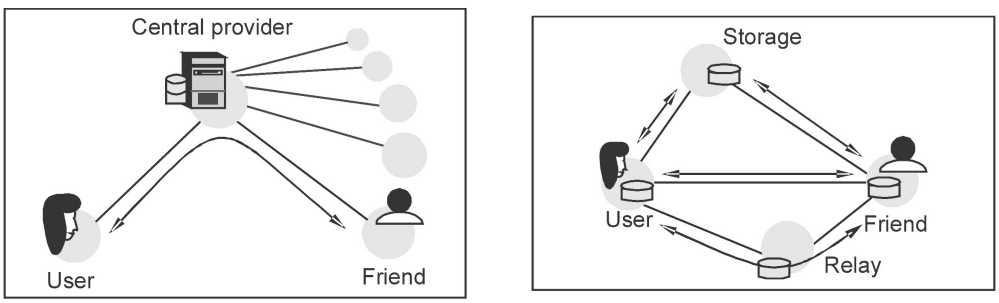
3.7.2 Tools for Interactive Analysis

- List of interactive analysis tools are : JUNG, Netminer, igraph, SONIVIS, Commetrix, NetworkWorkbench, visone, CFinder etc.
  1. igraph

- igraph is a free software package for creating and manipulating undirected and directed graphs. It includes implementations for classic graph theory problems like minimum spanning trees and network flow, and also implements algorithms for some recent network analysis methods, like community structure search.
- igraph is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. igraph is open source and free. igraph can be programmed in R, Python and C/C++.
- Igraph uses space and time efficient data structures and implements the current state-of-the-art algorithms.
- The igraph library has a layered architecture, the three layers are connected through well defined interfaces. Each layer can be replaced with an alternate implementation without changing the other components.

3.8 Decentralized Online Social Networks

- Current Online Social Networks (OSN) are web services run on logically centralized infrastructure. Large OSN sites use content distribution networks and thus distribute some of the load by caching for performance reasons; nevertheless there is a central repository for user and application data.
- This centralized nature of OSNs has several draw-backs including scalability, privacy, dependence on a provider, need for being online for every transaction, and a lack of locality.
- Fig. 3.8.1 shows logical communication flows in centralized OSN and decentralized OSN architectures.



(a) Logical communication flows in centralized OSN

(b) Logical communication flows in decentralized OSN

Fig. 3.8.1

- A decentralized online social network (DOSN) is a distributed system for social networking with no or limited dependency on any dedicated central infrastructure.
- In a decentralized social networking framework, a user does not need to join any particular social networking service such as Facebook or MySpace.

- Instead, the user chooses a server which he trusts to host his own data such as his FOAF (Friend-Of-A-Friend) file, his activity log and his photo albums. Given that we refer to these files with their URIs, they can actually be stored on different servers.
- Decentralized OSN can also be implemented with the use of **Peer-to-Peer (P2P)** network. A P2P network is a distributed network in which nodes are connected with each other to participate in processing, memory, and bandwidth intensive tasks.
- These networks scale better than centralized server architectures without the need of costly centralized resources.
- P2P networks have been popular mostly as file sharing networks (such as KaZaA, BitTorrent, etc.) and sometimes as collaborative sharing networks , but have not been used as a medium for online social networking. The inherent nature of peer-to-peer connection between users in a social network makes OSN suitable for peer-to-peer architecture.
- In file sharing P2P systems like Gnutella, most of the users are free riders. In contrast, P2P applications like Skype, where users tend to stay connected to the network to receive calls from their friends, shows the potential that P2P holds as implementation infrastructure for OSN.

3.8.1 Architecture of a Distributed Online Social Network

- Fig. 3.8.2 shows general architecture of a distributed online social network.

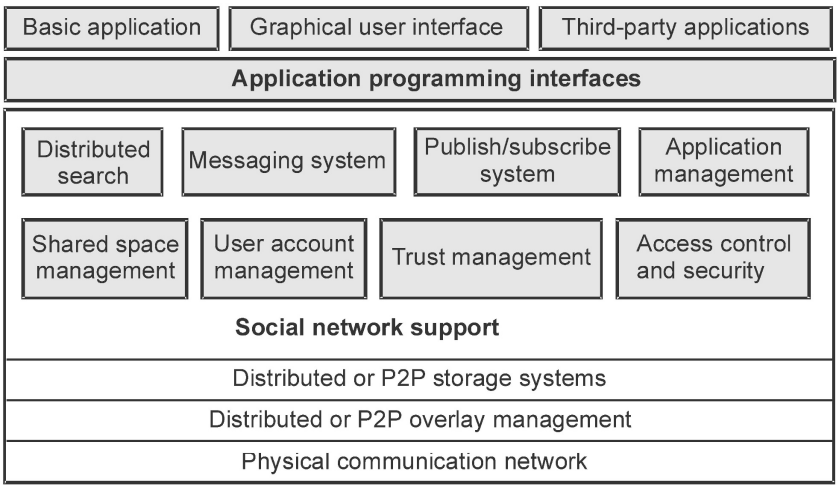


Fig. 3.8.2 : General architecture of a distributed online social network

- Physical communication network is lower layer in the architecture. It is the Internet or ad hoc network.
- The distributed or P2P overlay management : This layer provides core functionalities to manage resources in the supporting infrastructure of the system. This layer provides higher layers the capabilities of looking up resources, routing messages, and retrieving information reliably and effectively among nodes in the overlay.

- Decentralized data management layer : It provides functionalities of a distributed or peer-to-peer information system to query, insert, and update various persistent objects to the systems.
- The social networking layer implements all basic functionalities and features that are provided by contemporary centralized social networking services.
- The top layer of the architecture includes the user interface to the system and various applications built on top of the development platform provided by the DOSN.
- In current DOSNs, data are stored and maintained available at some peers. The data of decentralized OSNs is distributed across multiple administrative domains.
- Application servers run on desktop machines (i.e., peers) owned by users. In general, hosting personal data on peers is more privacy-preserving than delegating control to a third-party service provider.

3.8.2 Proposed DOSN Approaches

1. Safebook

- Safebook consists of three-tier architecture with a direct mapping of layers to the OSN levels. Fig. 3.8.3 shows safebook components.

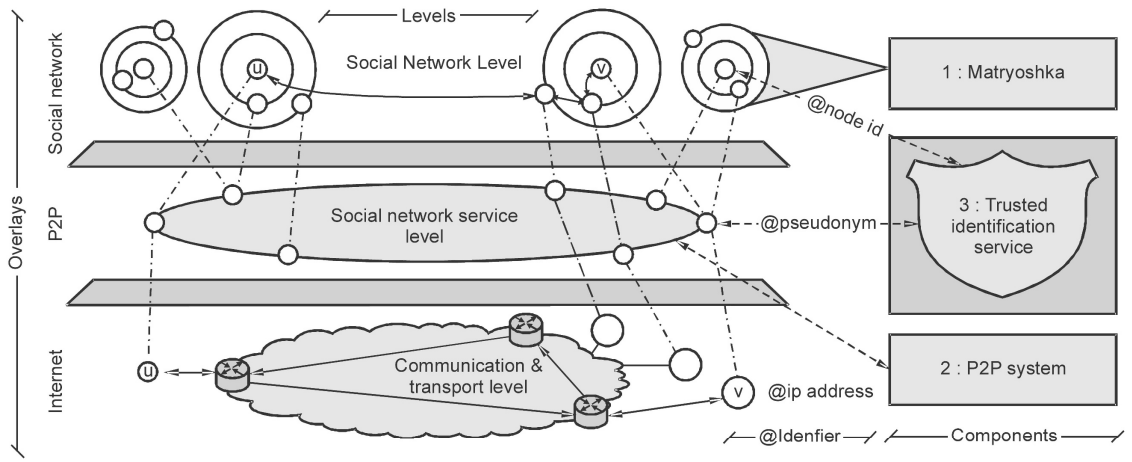


Fig. 3.8.3 Safebook components

1. The user-centered Social Network layer implementing the SN level of the OSN;
2. The Peer-to-Peer substrate implementing the SNS services;
3. The Internet, representing the CT level

- Each party in Safebook is thus represented by a node that is viewed as a host node in the Internet, a peer node in the P2P overlay, and a member in the SN layer.
- The nodes in Safebook form two types of overlays :
  1. Set of Matryoshkas, concentric structures in the SN layer providing data storage and communication privacy created around each node;
  2. P2P substrate, providing lookup services.

- In addition to these nodes, Safebook also features a Trusted Identification Service (TIS), providing each node unambiguous identifiers: the Node Identifier for the SN level and a Pseudonym.
- Matryoshkas are concentric rings of nodes built around each member's node in order to provide trusted data storage, profile data retrieval and communication obfuscation through indirection.
- Trusted Identification Service (TIS) assures that each Safebook user gets at most one unique identifier in each category of identifiers.
- Based on an out of band identification procedure, the TIS grants each user a unique pair of a node identifier and a pseudonym, computed as the result of a keyed hash function on the set of properties that uniquely identify a party in real life, such as full name, birth date, birth place and so on.
- Safebook implements different OSN operations :
  1. Account creation;
  2. Data publication;
  3. Data retrieval;
  4. Contact request and acceptance;
  5. Message management

➤ FOAF (Friend of a friend)

- FOAF is used for describing people profiles, their relationships and their activities online. FOAF aims to create a linked information system about people, groups, companies and other kinds of thing.
- If people publish information in FOAF document format, machines will be able to make use of that information.
- If those files contain “see also” references to other such documents in the Web, we will have a machine-friendly version of today's hypertext web. FOAF documents are usually represented in RDF.
- Users query and manage these profiles through open Web-based protocols such as WebDAV or SPARQL/Update

3.9 Multi-Relational Characterization of Dynamic Social Network Communities

- Dynamic social networks are social networks that take into account changes over time. They not only model relations between human beings in terms of interpersonal interactions, but also consider the evolution of these relations, i.e. the way and the extent by which they change over time.
- Dynamic social networks can be useful to model and analyze human relationships in several potential scenarios: the informal social relationships of individuals within a family or a group of friends; the structured collaboration of employees in a large enterprise
- **Mutual awareness** refers to a relationship developed through observable interactions between two people. We can define mutual awareness computationally by contextual use of links in social media.

- If Rupali, comments on Rakshita's blog post, Rakshita is aware of Rupali, but Rupali cannot be certain that Rakshita is aware of her, if her comment is unread. Subsequently, if Rakshita comments on Rupali's blog post, there is mutual awareness between the two.
- Mutual awareness can be asymmetric; the asymmetry can arise, for example, when one person is a celebrity, or is touch with more people than the other. In addition, mutual awareness strength can **change over time**.
- **Transitive awareness** refers to a relationship, computed via a mutual awareness measure between two connected people on a network. We can compute transitive awareness between a connected pair of users on a social network graph, through mutual-awareness expansion. We can use a random walk based distance, with an efficient method for mutual awareness expansion, to extract communities.
- Weblogs (blogs) adapt web technology to allow for instant, updated and frequent communication of information such as events, personal interests, thoughts, and news.
- The conversational nature of blogs can be used to uniquely identify topics and similar areas of interest. These similarities can then be used to build (virtual) community, and specifically identify virtual community, in blogs.
- Blogs are an excellent example of social hypertext because the comments from blogs are hyperlinks to other sources on the web and to the commenter's blog. Since the comments are associated with a particular blog's post and are in chronological order, blogs can facilitate members' social interactions.
- Blog communities are different from traditional web communities
- Properties of blogs
  1. Temporal dynamics : Blogs represent easily editable content.
  2. Event Locality : A typical blog entry is time sensitive.
  3. Link Semantics : A hyperlink can have different semantics.
  4. Community Centric : People that interested in each others' content

➤ Extracting Communities Based on Mutual Awareness Structure

- Extracting blog communities by using following methods :
  1. Computing Mutual Awareness
  2. Ranking-Based Clustering Method
- Mutual awareness is affected by
  1. Type of action
  2. Number of actions for each type
  3. When the action occurred
- Mutual awareness depends on sustained actions. For each action type k at time t, compute Temporal action matrix $X_{k,t}$.
- Each entry $x_{ij,k,t}$ of matrix represents the number of times the k^{th} action a_k was performed by blogger i on blogger j. For example : Blogger i leaves a comment on blogger j's entry.

- Effect of actions to mutual awareness diminish gradually

$$X_k = \sum_{t=t_0}^{T} X_{k,t}\, e^{-\lambda_k (T-t)}$$

Where λ_k is decaying factor for the action type k

- Fig. 3.9.1 shows framework for measuring and evaluating community in blogs.

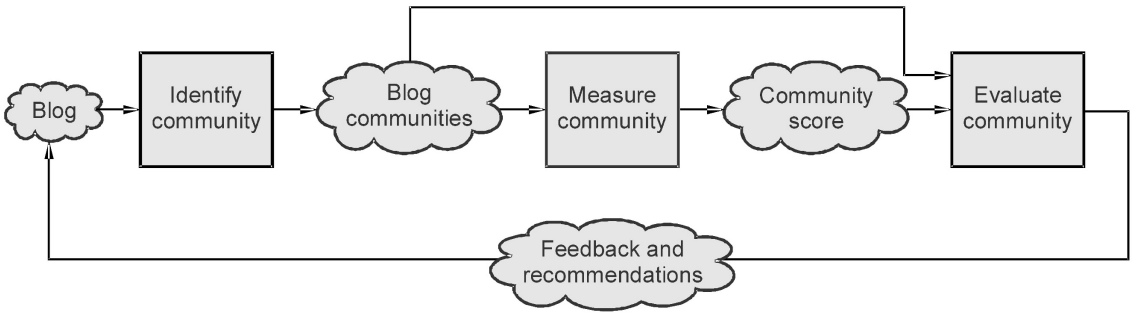


Fig. 3.9.1 : Framework for measuring and evaluating community in blogs

- Framework consists of three processes.
  1. The first main step in this framework is to identify community in the blogs, typically represented not by a single community, but as a set of blog communities. Visualization of network structure is one way to accomplish this step.
  2. Measure community in the set of blog communities by creating a community score.
  3. Compare the community score for the blog and the set of blog communities with others in order to evaluate community. The resulting feedback and recommendations from evaluating the community may then be used to identify strategies for further improving the amount of community in the blogs.

3.10 Questions with Answers

3.10.1 Two Marks Questions with Answers

Q. 1 Define clique.

Ans. : A clique in a graph is a **maximal complete** subgraph of three or more nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.

Q. 2 Define n-clique.

Ans. : An n-clique is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than n.

Q. 3 What is dendrogram ?

Ans. : Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

Q. 4 What do you mean community detection ?

Ans. : Community detection : Discovering groups in a network where individuals' **group memberships** are not explicitly given

Q. 5 What is web community ?

Ans. : Web community is a set of web objects (documents and users) that includes its own semantic and logical structures. Given a graph, a community is defined as a set of nodes that are more densely connected to each other than to the rest of the network nodes.

Q. 6 What is FOAF ?

Ans. : FOAF is used for describing people profiles, their relationships and their activities online. FOAF aims to create a linked information system about people, groups, companies and other kinds of thing. If people publish information in FOAF document format, machines will be able to make use of that information.

Q. 7 Explain decentralized online social network.

Ans. : A decentralized online social network (DOSN) is a distributed system for social networking with no or limited dependency on any dedicated central infrastructure

Q. 8 What is use of Igraph tool ?

Ans. : igraph is a free software package for creating and manipulating undirected and directed graphs. It includes implementations for classic graph theory problems like minimum spanning trees and network flow, and also implements algorithms for some recent network analysis methods, like community structure search

Q. 9 Explain difference between communities and social networks.

Ans. :

Communities	Social Networks
Communities are held together by some common interests of a large group of people. Although there may be pre-existing interpersonal relationship between members of a community, it is not required. So new members usually do not know most of the people in the community.	Social Networks are held together by pre-established interpersonal relationships between individuals. So you know everyone that is directly connected to you.
Any one person may be part of many communities.	Each person has one social network. But a person can have different social graphs depending on what relationship we want to focus on
They have overlapping and nested structure	They have a network structure.

Q. 10 Explain dynamic social network.

Ans. : Dynamic social networks are social networks that take into account changes over time. They not only model relations between human beings in terms of interpersonal interactions, but also consider the evolution of these relations, i.e. the way and the extent by which they change over time

Q. 11 What is mutual awareness ?

Ans. : Mutual awareness refers to a relationship developed through observable interactions between two people. We can define mutual awareness computationally by contextual use of links in social media.

3.10.2 Fill in the Blanks

Q. 1 HITS stands for ----------.

Q. 2 ------- is one or a set of Web pages that provides collections of links to authorities

Q. 3 Decompose data objects into a several levels of nested partitioning (tree of clusters), called a ----------.

Q. 4 --------- communities are different from traditional web communities

Q. 5 Blogs are often considered as ----------.

Q. 6 Quality function supports for finding good partitions. The most popular quality function is the ----------.

Q. 7 ---------- awareness refers to a relationship, computed via a mutual awareness measure between two connected people on a network

Q. 8 A ------- is a page with a list of hyperlinks to valuable pages on the topic, i.e. points to many good authorities.

Q. 9 The ---------- axis of the dendrogram represents the distance or dissimilarity between clusters

Q. 10 ---------- algorithms are highly effective at discovering community structure in both computer-generated and real-world network data.

Q. 11 -------- awareness refers to a relationship developed through observable interactions between two people

Q. 12 A clique in a graph is a ---------- subgraph of three or more nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.

3.10.3 Multiple Choice Questions

Q. 1 An ---------- is a page with good contents on a topic, and is pointed to by many good hub pages.

(a) Hub (b) Authority (c) HITS (d) All of these

Q. 2 HITS stands for ----------

(a) Hypertext Induced Topics Search

(b) Hypertext Induced Topics System

(c) Hub Induced Topics Search

(d) Hypertext Included Topics Search

Q. 3 Blog rolls is lists of discussion partners on a -------- blog.

(a) Group (b) personal (c) welfare (d) none of these

Q. 4 Ontology ---------- creates links between two original ontologies.

(a) Graph (b) ontomorph (c) alignment (d) none of these

Q. 5 Decompose data objects into a several levels of nested partitioning, called ----------.

(a) dendrogram (b) binary tree (c) decision tree (d) graph

Q. 6 The methods for detecting communities are roughly classified into the following categories:

(a) divisive algorithms (b) modularity optimization

(c) spectral algorithms (d) All of these

Q. 7 Pages which are not very relevant but point to pages in the Root are called ----------

(a) Authority (b) HITS (c) Hubs (d) None of these

➤ Answers of Fill in the Blanks

1.	hyperlink-induced topic search	7.	Transitive
2.	hub	8.	hub
3.	dendrogram	9.	horizontal
4.	Blog	10.	Newman-girvan
5.	personal publishing or digital diary	11.	Mutual
6.	modularity	12.	maximal complete

➤ Answers of Multiple Choice Questions

1.	b	2.	a	3.	b	4.	c	5.	a	6.	d	7.	c

❑❑❑

4 Predicting Human Behaviour and Privacy Issues

Scope of the Syllabus

Understanding and predicting human behaviour for social communities - User data management - Inference and Distribution - Enabling new human experiences - Reality mining - Context - Awareness - Privacy in online social networks - Trust in online environment - Trust models based on subjective logic - Trust network analysis - Trust transitivity analysis - Combining trust and reputation - Trust derivation based on trust comparisons - Attack spectrum and countermeasures.

4.1 Understanding and Predicting Human Behaviour for Social Communities

- Currently, online social networks such as Facebook, Twitter, Google+, LinkedIn, have become extremely popular all over the world and play a significant role in people's daily lives. People access online social networks using both traditional desktop PCs and smartphone.
- Online social network user behavior covers various social activities that users can do online, such as friendship creation, content publishing, profile browsing, messaging and commenting.
- According to Technorati, about 75,000 new blogs and 1.2 million new posts giving opinion on products and services are generated every day. Also massive data generated every minute on common social network sites.
- Users opinions on social network sites can be referred to as discovery and recognition of positive or negative expression on diverse subject matters of interest.

4.1.1 User Data Management, Inference and Distribution

- Social information is also leveraged in conjunction with location and collocation data in mobile applications such as Loopt, Foursquare and Google's Latitude. These applications collect, store and use sensitive social information.

- The state of the art is to collect and manage such information within an application, thus offering social functionalities limited only to the context of the application, as in the examples above or to expose this information from platforms that specifically collect it, such as online social networks (OSNs).
- For example, Facebook and Google allow 3rd-party application developers and websites to access the social information of millions of users stored in their OSNs.
- In order to apply user information across a range of services and devices, there is a need for standardization of user related data and the architecture that enables their interoperability.
- Being a logically centralized data storage, it can be mapped to physically distributed configurations and should allow data to be accessed in a standard format.
- To manage social network data, systems either use a relational engine, which comes with convenient features such as indexes and transactional support or use a native graph engine.
- Publication of linked data is currently booming on the web, fostered by important governmental agencies, companies and scientists from various domains. Unrelated communities of users are increasingly interested in bulk-exporting, querying and integrating heterogeneous data across the web using this formalism. Database administrators either use legacy relational systems or native triple stores to manage linked data.

4.2 Enabling New Human Experiences

- Extensive information gives rise to challenge of automatic summarization. Opinion definition and opinion summarization are essential techniques for recognizing opinion.
- Opinion definition can be located in a text, sentence or topic in a document; it can also reside in the entire document.
- Opinion summarization sums up different opinions aired on piece of writing by analyzing the sentiment polarities, degree and the associated occurrences.

4.2.1 Reality Mining

- One of the most important applications of reality mining may be the automatic mapping of social networks. Reality mining is defined as the study of human social behavior based on wireless mobile phone sensed data.
- To overcome the discrepancy between online and “offline” networks, reality mining techniques can be empowered to approximate both worlds, proving awareness about people actual behavior.
- Understanding the social behavior patterns of different subpopulations and the mixing between them is critical to the delivery of public health services, because different subpopulations have different risk profiles and different attitudes about health-related choices. The use of reality mining to discover these social behavior patterns can potentially provide great improvements in health education efforts and behavioral interventions.
- Fig. 4.2.1 shows a smart phone is programmed to sense other people using Bluetooth. In figure, you see a smart phone that is programmed to sense and report continuously on its user's location, who else is nearby, the user's call and SMS patterns and how the user is moving.

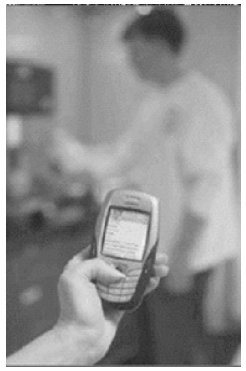


Fig. 4.2.1 : Smart phone programmed to sense other people using bluetooth

- Careful analysis of these data shows different patterns of behavior depending upon the social relationship between people. Fig. 4.2.2 shows the patterns of proximity among the participants during one day; even casual examination shows that the students were part of two separate groups.

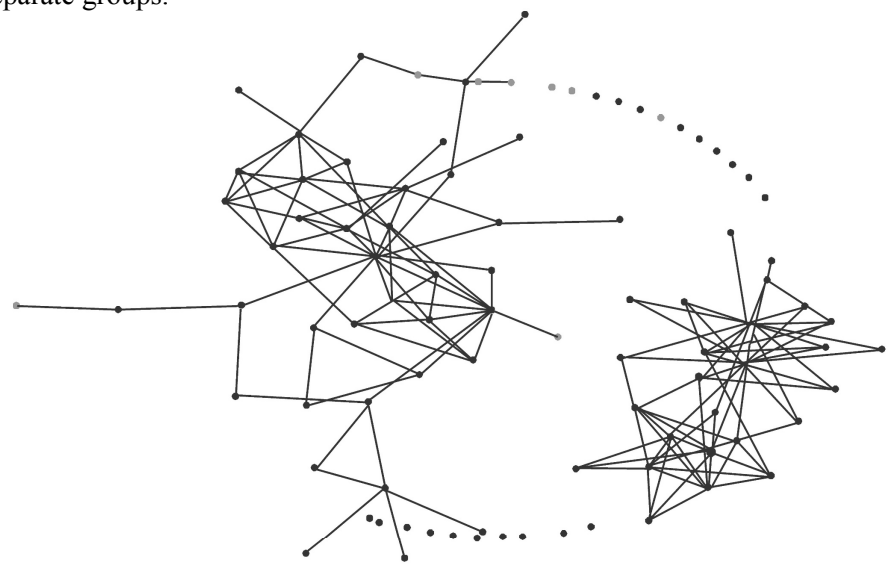


Fig. 4.2.2 : Pattern of proximity between people during one day

- It typically analyzes sensor data from mobiles to extract subtle patterns that help to predict and understand future human behavior.

4.2.2 Context-Awareness

- Context is a combination of any information that can be sensed or received by an entity which is useful to catch events and situations.
- Context-aware computing uses information about an end user's or object's environment, activities, connections and preferences to improve the quality of interaction with that end user or object.

- A contextually aware system anticipates the user's needs and proactively serves up the most appropriate and customized content, product or service.
- Applications that use context, whether on a desktop or in a mobile or ubiquitous computing environment, are called context-aware.
  1. Context is available, meaningful and carries rich information in such environments.
  2. That users' expectations and user experience is directly related to context, acquiring, representing, providing and using context becomes a crucial enabling technology for the vision of disappearing computers in everyday environments.

4.3 Privacy in Online Social Networks

- Online Social Networks (OSNs) have become part of daily life for millions of users. Users building explicit networks that represent their social relationships and often share a wealth of personal information to their own benefit.
- The potential privacy risks of such behavior are often underestimated or ignored. The problem is exacerbated by lacking experience and awareness in users, as well as poorly designed tools for privacy-management on the part of the OSN.

➤ Definition of an online social network

- An **online social network** is a web-based service that allows individuals to :
  1. Construct a public or semi-public profile within the service,
  2. Articulate a list of other users with whom they share a connection,
  3. View and traverse their list of connections and those made by others within the service.
- "**UseNet Newsgroups**" is the first online social networks and developed by Duke University graduate students Tom Truscott and Jim Ellis in 1979. Day by day, online social network is increasing in size and numbers.
- Online social networking has gained tremendous popularity amongst the masses. It is usual for the users of Online Social Networks (OSNs) to share information with friends however they lose privacy.
- Privacy has become an important concern in online social networks. Users are unaware of the privacy risks involved when they share their sensitive information in the network.
- One of the fundamental challenging issues is measurement of privacy. It is hard for social networking sites and users to make and adjust privacy settings to protect privacy without practical and effective-way to quantify, measure and evaluate privacy.
- Fig 4.3.1 shows online social network model.

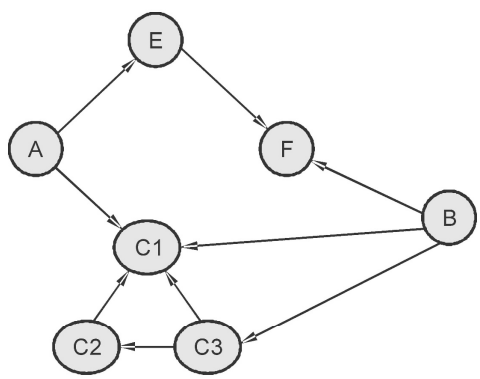


Fig. 4.3.1 : OSN model

- The principle importance of an OSN is to make relationships with different users and accomplish such relationships for allocating resources of different nature. So, it is acknowledged that any access control model for OSNs ought to be relation-ship based.
- Facebook is a popular online social network. A user's Facebook profile contains a wealth of personal information, including name, photo, date of birth, contact information, sexual orientation and relationship status, political and religious views, personal interests, hobbies, education history and more.
- This information is made available to members of the user's social network, allowing friends to stay in touch and up-to-date with each other's lives. At the same time, Facebook generates revenue by targeting ads to highly specific demographics.

4.3.1 Trust in Online Environment

- Researchers from different sub-disciplines in computer science have tackled some of the problems that arise in OSNs and proposed a diverse range of "privacy solutions". These include software tools and design principles to address OSN privacy issues.
- Evaluation trust can be interpreted as the reliability of something or somebody and the decision trust captures broader concept of trust.
- Evaluation trust : Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends.
- Decision trust : Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible.
- Many social network platforms have been developed on the Web such as Twitter and Facebook. In those networks it may be the case that a lot of the end-users are usually physically unknown with each other.
- In this case if two unknown participants wish to communicate with each other for various reasons, the evaluation of their trustworthiness along a certain trust path between them within the social network is mandatory. But the level of trustworthiness may vary and it is sometimes subjective and depends on the person's specific role within the network.
- It is not an easy task as trust cannot easily be defined through mathematical formulas and algorithmic procedures. Trust may rely on several factors from psychological and sociological factors to computer security factors.
- The main difference between users and the service provider is the type of information they can access. A user or outsider can generally only view public information. The service provider can generally view all data in the system, including private uploads, browsing behavior, IP addresses, etc.
- It is also the service provider who decides which data is stored, how long it is kept and how it is used or distributed. The user is also dependent on the service provider for tools to protect his privacy. Therefore, trust plays a big role in the relationship between a user and the service provider.

- General properties of trust are as follows :
  1. Trust is a measurable belief
  2. Trust is directed
  3. Trust exists in time
  4. Trust evolves in time, even within the same transaction
  5. Trust is a subjective belie

4.3.2 Trust Models based on Subjective Logic

- Subjective logic is a type of probabilistic logic that explicitly takes uncertainty and belief ownership into account. In general, subjective logic is suitable for modeling and analysing situations involving uncertainty and incomplete knowledge.
- Subjective logic represents a practical belief calculus that can be used for calculative analysis trust networks.
- Arguments in subjective logic are subjective opinions about states in a state space. A binomial opinion applies to a single proposition and can be represented as a Beta distribution.
- A multinomial opinion applies to a collection of propositions and can be represented as a Dirichlet distribution. Fig. 4.3.2 shows subjective logic.

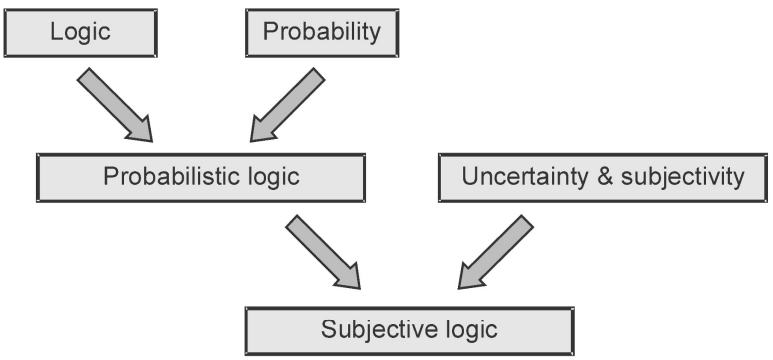


Fig. 4.3.2 Subjective logic

- Trust transitivity means, for example, that if Alice trusts Bob who trusts Eric, then Alice will also trust Eric. This assumes that Bob actually tells Alice that he trusts Eric, which is called a recommendation.
- It can be shown that trust is not always transitive in real life. For example : The fact that Alice trusts Bob to look after her child, and Bob trusts Eric to fix his car, does not imply that Alice trusts Eric for looking after her child or for fixing her car .
- However, under certain semantic constraints, trust can be transitive and a trust system can be used to derive trust. In the last example, trust transitivity collapses because the scopes of Alice's and Bob's trust are different.

- We define trust scope as the specific type(s) of trust assumed in a given trust relationship. In other words, the trusted party is relied upon to have certain qualities and the scope is what the trusting party assumes those qualities to be.
- Let us assume that Alice needs to have her car serviced, so she asks Bob for his advice about where to find a good car mechanic in the city. Bob is thus trusted by Alice to know about a good car mechanic and to tell his honest opinion about that.
- Bob in turn trusts Eric to be a good car mechanic. This situation is illustrated in Fig. 4.3.3, where the indexes indicate the order in which the trust relationships and recommendations are formed.

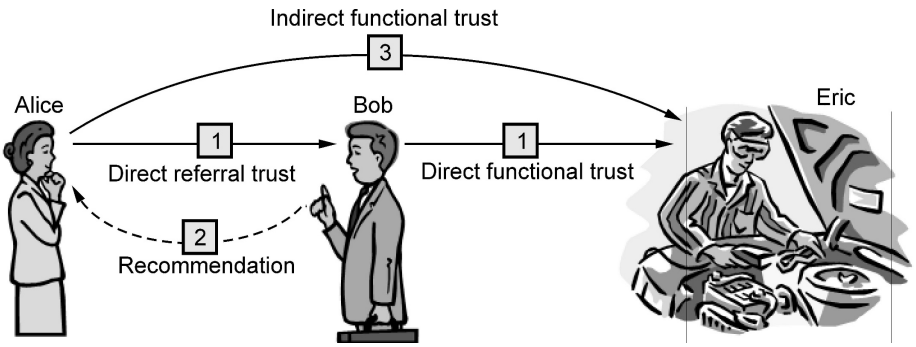


Fig. 4.3.3 : Transitive trust principle

- It is important to separate between trust in the ability to recommend a good car mechanic which represents referral trust and trust in actually being a good car mechanic which represents functional trust.
- A single trust relationship can be expressed as a directed arc between two nodes that represent the trust source and the trust target of that arc. For example the arc [A; B] means that A trusts B.
- The symbol ““:”“ will be used to denote the transitive connection of two consecutive trust arcs to form a transitive trust path. Fig. 4.3.4 shows transitive serial combination of trust arcs.

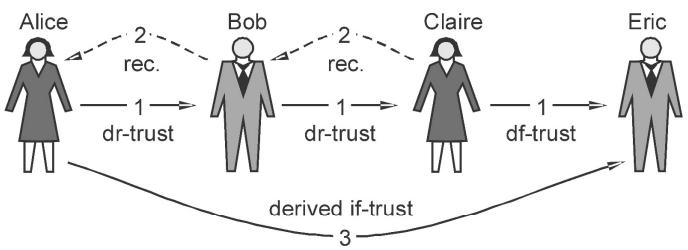


Fig. 4.3.4 : Transitive serial combination of trust arcs

- The trust relationships of above Fig. 4.3.4 can be expressed as :

  ([A,E]) = ([A,B] : [B,C] : [C,E])

  where the trust scope is implicit.

➤ Parallel trust combination

- It is common to collect advice from several sources in order to be better informed when making decisions. This can be modelled as parallel trust combination shown in Fig. 4.3.5, where again the indexes indicate the order in which the trust relationships and recommendations are formed.

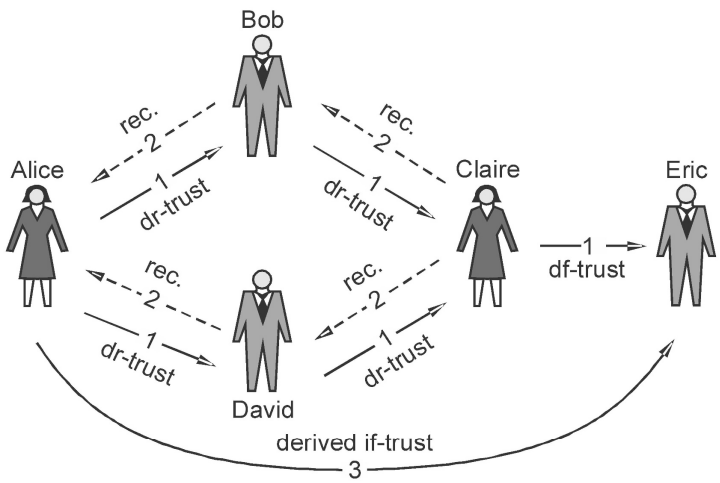
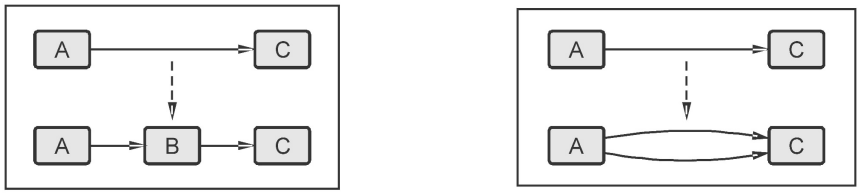


Fig. 4.3.5 : Parallel combination of trust paths

- We will use the symbol ◊, to denote the graph connector for parallel combination of trust paths. The ◊, symbol visually resembles a simple graph of two parallel paths between a pair of agents.
- Alice's combination of the two parallel trust paths from her to Eric is then expressed as :

$$([A,E]) = ((([A,B] : [B,C] \lozenge ([A,D] : [D,C])) : [C,E])$$

- Simplification of a trust network consists of including as many arcs as possible from the original trust network, while still maintaining a canonical expression. Graphs that can be represented as canonical expressions with our structured notation are known as directed series-parallel graphs **(DSPG)**.
- Definition of directed series-parallel composition : A directed series composition consists of replacing an arc [A; C] with two arcs [A; B] take [B; C] where B is a new node.
- A directed parallel composition consists of replacing an arc [A; C] with two arcs $[A; C]_1$ and $[A; C]_2$.
- When implementing the serial ":"as binary logic "AND" and the parallel "◊" as binary logic "OR", the results would be equal. However, when implementing ":" and "◊" as probabilistic multiplication and co-multiplication respectively, the results would be different.
- The principle of directed series and parallel composition are shows in Fig. 4.3.6.



(a) Series graph composition **(b) Parallel graph composition**

Fig. 4.3.6

4.4 Trust Network Analysis

- Trust networks consist of transitive trust relationships between people, organizations and software agents connected through a medium for communication and interaction.
- Trust is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action and in a context in which it affects his own action.
- If there is a notion of trust, there is also the notion of distrust as the opposite of trust. To distrust is not the same as having a lack of knowledge whether to trust or not, it is to know not to trust.
- To have a lack of knowledge whether to trust or distrust is ignorance. Trust or distrust is extremes of a continuous scale. The scale depends on the trust model and the distance of the value from the actual extremes is called uncertainty.
- A method for trust net-work analysis using subjective logic (TNA-SL) has been described by Jøsang et al.
- Subjective logic is a trust algebra based on bayesian theory and Boolean logic. It presents an opinion as three interdependent values.
- Trust Network Analysis with Subjective Logic (TNA-SL) is based on analyzing trust networks as directed series-parallel graphs that can be represented as canonical expressions, combined with measuring and computing trust using subjective logic.
- One advantage of TNA-SL is that negative trust can be explicitly expressed and propagated. In order for distrust to be propagated in a transitive fashion, all intermediate referral arcs must express positive trust, with only the last functional arc expressing negative trust.
- The main disadvantage of TNA-SL is that a complex and cyclic network must be simplified before it can be analyzed, which can lead to loss of information. While the simplification of large highly connected networks could be slow, heuristic techniques can significantly reduce the computational effort.
- By Gambetta and belief is thus defined as; the subjective probability that a particular action will be preformed, before it can be monitored or independently of the ability to ever monitor it.

4.4.1 Operators for Deriving Trust

- Beliefs are subjective, that is they do not contain truth in a traditional sense, they only contain a subjective or experienced probability and they focus on what we have evidence to support not what the actual outcome is.
- A belief is held by an agent. The term **opinion** is used to denote such subjective beliefs held by an agent. An opinion is not shared by agents, they might have the same amount of belief in the same thing but they each have their own opinion. Fig. 4.4.1 shows opinion triangle.

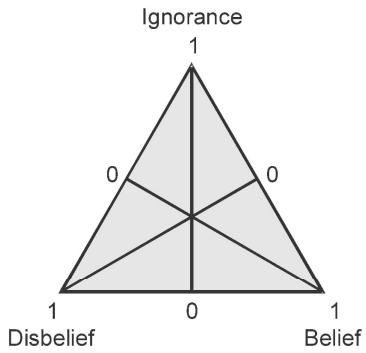


Fig. 4.4.1 : Opinion triangle

- The set of all possible opinions is denoted as Ω. Thus, any given opinion ω can be presented graphically as a point within the opinion triangle.
- In subjective logic, beliefs are represented on binary state spaces, where each of the two possible states can consist of sub-states. Belief functions on binary state spaces are called subjective opinions.
- Subjective logic represents a specific belief calculus that uses a belief metric called opinion to express beliefs. An opinion denoted by ω_x^A = (b, d, u, a) expresses the relying party A' s belief in the truth of statement "X".
- Here b, d and u represent belief, disbelief and uncertainty respectively, where b, d, u, $\in$ [0, 1] and b + d + u = 1 .
- The confidence parameter can be define a equal to (1 - c), i.e. the confidence of a trust value is equivalent to the certainty of the corresponding opinion.
- The parameter a $\in$ [0,1] is called the base rate and is used for computing an opinion's probability expectation value that can be determined as $E\left(\omega_x^A\right)$ = b + au.
- Transitivity is used to compute trust along a chain of trust edges.
- Cumulative fusion is equivalent to Bayesian updating in statistics. The cumulative fusion of two possibly conflicting opinions is an opinion that reflects both opinions in a fair and equal way.

4.5 Trust Transitivity Analysis

- In order to compute more reasonable trust value between two unknown participants, a critical and challenging problem is to make clear how and to what extent trust is transitive along a social trust path.
- If there is a trust path linking two nonadjacent participants, the source participant can evaluate the trustworthiness of the target one along an existing path based on the trust transitivity property.
- The path with trust information linking the source participant and the target one is called **a social trust path**.

- Trust transitivity, as trust itself, is a human mental phenomenon, so there is no such thing as objective transitivity, and trust transitivity therefore lends itself to different interpretations.

4.6 Combining Trust and Reputation

- Online trust and reputation systems are emerging as important decision support tools for selecting online services and for assessing the risk of accessing them.
- Trust and reputation is a security mechanism in environments where several entities communicate and interact. This security mechanism is based on two attributes found in human relationships : Trust and reputation.
- There are several approaches to represent trust on which agents base their decisions. Most of them use a numerical representation of several trust states. Marsh for example represent trust as a continuous variable in a defined interval where certain subintervals imply how much an entity is trusted.
- Reputation in general is an estimation how an agent will behave in the future based on observations of its past behaviour.
- Reputation can either be the accumulation of several observations from different communicating agents or it can be based only on the experience a single agent has made in the past.
- Reputation is used because it offers an additional source for agents to rely on when making trust decisions. This source is necessary because it is hardly possible for an agent to consider every aspect when making a trust decision.
- In addition reputation can consist of experiences from several agents as mentioned before which can be an advantage because those accumulated experiences offer information a single agent could not obtain.
- A general characteristic of reputation systems is that they provide global reputation scores, meaning that all the members in a community will see the same reputation score for a particular agent. On the other hand, trust systems can in general be used to derive local and subjective measures of trust, meaning that different agents can derive different trust in the same entity.
- Another characteristic of trust systems is that they can analyse multiple hops of trust transitivity. Reputation systems on the other hand normally compute scores based on direct input from members in the community which is not based on transitivity. Still there are systems that have characteristics of being both a reputation system and a trust system. The matrix below shows examples of the possible combinations of local and global scores and trust transitivity or not.
- Reputation systems collect ratings about users or service providers from members in a community. The reputation centre is then able to compute and publish reputation scores about those users and services.

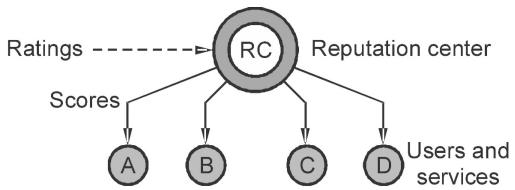


Fig. 4.6.1 : Reputation centre

- Fig. 4.6.1 shows a reputation centre where the dotted arrow indicate ratings and the solid arrows indicate reputation scores about the users.
- The compatibility between Bayesian reputation systems and subjective logic makes this a very flexible framework for analysing trust in a network consisting of both reputation scores and private trust values.

4.7 Trust Derivation Based on Trust Comparisons

- It is possible that different agents have different trust in the same entity, which intuitively could affect the mutual trust between the two agents. Fig. 4.7.1 shows deriving trust from conflicting trust.
- Two agents having similar point estimates about the same agent or proposition should induce mutual trust and dissimilar point estimates should induce mutual distrust.

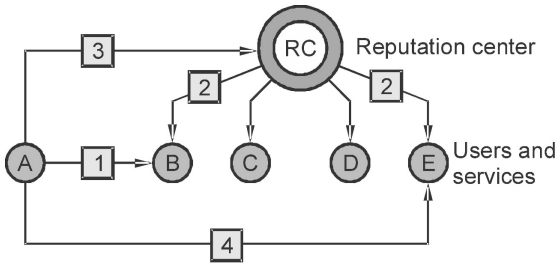


Fig. 4.7.1 : Deriving trust from conflicting trust

4.8 Attack Spectrum and Countermeasures

- Security objectives are requirements that have to be satisfied in order to protect the system from potential threats and attacks.
- The diversity of available OSN platforms opens doors for a variety of attacks on privacy of the users, integrity of their profiles and the availability of the user provided contents.

➤ 1. Plain impersonation

- With plain impersonation attack the adversary aims to create fake profiles for real-world users.
- The success of this attack strongly depends on the authentication mechanisms deployed in the registration process. Since many OSNs tend to authenticate email addresses by requesting confirmations for the registration emails, this attack can be easily performed if an email address is created in advance.

➤ 2. Profile cloning

- Aim is to create a profile for some user that is already in possession of some valid profile in the same network.
- This attack can be realized through the registration of the new profile using the same or similar content as the existing one.
- This is feasible in most OSN platforms since each profile is associated with some unique administrative id and an email address used during the registration.

- Many users hide their email address so that OSN users would not be able to distinguish between the original profiles and their clones registered with other email addresses.
- The scripted profile cloning is possible in Facebook, XING and the German sites StudiVZ and MeinVZ. In all these services but XING, CAPTCHAs were employed and CAPTCHA-breaking techniques were required. In the case of LinkedIn CAPTCHA mechanisms are not in place.

➤ 3. Profile hijacking attack

- Profile hijacking attack is to obtain control over some existing profile within an OSN platform. Normally password is used to protect the profile.
- Once hacker knows the password, then profile hacking is possible. Profiles are harvested from social-network specific queries using each network's search mechanism that contain these terms and the user's real name.
- The user is initially prompted for his real name, valid email address and a password. This suffices for creating a provisionary account in the service, which needs to be verified by accessing a private URL, sent to the user via email and entering the account's password.
- Receiving such messages and completing the verification process is trivial to be scripted and therefore can be carried out without human intervention

➤ 4. Profiles and sybil attacks

- Sybil attacks focus on creating multiple online user identities (Sybil identities) and try to achieve malicious results through these identities.
- In an Identity Clone attack (also called Profile Cloning attack), an adversary first creates similar or even identical profiles to impersonate victims in an OSN system. He then distorts the reputation and the value of a resource through the network involving faked profiles.
- Sybil attacks and identity clone attacks look somehow similar in appearance since both attacks need to create a number of online identities and use these identities to compromise the reputation and evaluation mechanisms in OSN systems.
- A Sybil attack can be used to affect the popularity, reputation, value and other characteristics of resources in OSN systems by using Sybil nodes. An adversary can boost invaluable resources and resource providers who have bad reputations.

➤ 5. Crawling and harvesting

- Crawling : To collect and aggregate publicly available information across multiple OSN profiles and applications in an automated way.
- Attacker simultaneously crawls across different OSN platforms are called **harvesting**.

➤ 6. Image retrieval and analysis

An automated attack aiming to collect multimedia information available with the OSN platform. This attack is typically followed by the subsequent analysis via automated pattern recognition tools to find links to the OSN profiles of displayed users.

4.9 Question with Answers

4.9.1 Two Marks Question with Answers

Q. 1 What is reality mining ?

Ans. : Reality mining is defined as the study of human social behavior based on wireless mobile phone sensed data.

Q. 2 Define crawling.

Ans. : To collect and aggregate publicly available information across multiple OSN profiles and applications in an automated way.

Q. 3 Define harvesting.

Ans. : Attacker simultaneously crawls across different OSN platforms are called harvesting.

Q. 4 What is sybil attacks ?

Ans. : Sybil attacks focus on creating multiple online user identities (Sybil identities) and try to achieve malicious results through these identities.

Q. 5 What is profile hijacking attack ?

Ans. : Profile hijacking attack is to obtain control over some existing profile within an OSN platform. Normally password is used to protect the profile.

Q. 6 What is use of reputation systems ?

Ans. : Reputation systems collect ratings about users or service providers from members in a community. The reputation centre is then able to compute and publish reputation scores about those users and services.

Q. 7 What is TNA-SL ?

Ans. : TNA-SL takes directed trust edges between pairs as input and can be used to derive a level of trust between arbitrary parties that are interconnected through the network.

Q. 8 What is trust network ?

Ans. : Trust networks consist of transitive trust relationships between people, organizations and software agents connected through a medium for communication and interaction.

Q. 9 Define online social network.

Ans. : An **online social network** is a web-based service that allows individuals to :

1. Construct a public or semi-public profile within the service, 2. Articulate a list of other users with whom they share a connection.

View and traverse their list of connections and those made by others within the service.

4.9.2 Fill in the Blanks

Q. 1 An online social network is a ---------- service that allows individuals to construct a public or semi-public profile within the service.

Q. 2 Subjective logic is a type of --------- logic that explicitly takes uncertainty and belief ownership into account.

Q. 3 A ---------- relationship can be expressed as a directed arc between two nodes that represent the trust source and the trust target of that arc.

Q. 4 Subjective logic is a -------- based on Bayesian theory and Boolean logic

Q. 5 A belief is held by an agent. The term -------- is used to denote such subjective beliefs held by an agent

Q. 6 The path with trust information linking the source participant and the target one is called a social ----------.

Q. 7 ---------- attack is to obtain control over some existing profile within an OSN platform

Q. 8 ----- attacks focus on creating multiple online user identities (sybil identities) and try to achieve malicious results through these identities.

Q. 9 In subjective logic, --------- are represented on binary state spaces, where each of the two possible states can consist of sub-states.

Q. 10 ---------- attack is to obtain control over some existing profile within an OSN platform.

Q. 11 Attacker simultaneously crawls across different OSN platforms are called ----------.

Q. 12 ---------- Mining is defined as the study of human social behavior based on wireless mobile phone sensed data.

Q. 13 ---------- in general is an estimation how an agent will behave in the future based on observations of its past behavior.

4.9.3 Multiple Choice Questions

Q. 1 TNA-SL stands for ----------

(a) Trust Network Analysis with subjective logic

(b) Trust Node Analysis with subjective logic

(c) Trust Network Area with subjective logic

(d) Trust Network Analysis with same logic

Q. 2 Set of all possible opinions is denoted as ---------

(a) Θ (b) Ω (c) α (d) ω

Q. 3 Attacker simultaneously crawls across different OSN platforms are called ----------.

(a) Crawling (b) Sybil (c) harvesting (d) Profile cloning

Q. 4 ---------- in general is an estimation how an agent will behave in the future based on observations of its past behavior.

(a) Profile (b) Reputation (c) trust (d) None of these

Q. 5 An online social network is a web-based service that allows individuals to:

(a) construct a public or semi-public profile within the service,

(b) articulate a list of other users with whom they share a connection,

(c) view and traverse their list of connections and those made by others within the service.

(d) All of these

➤ Answers of Fill in the Blanks

1.	web based	8.	Sybil
2.	probabilistic	9.	beliefs
3.	single trust	10.	Profile hijacking
4.	trust algebra	11.	harvesting
5.	opinion	12.	Reality
6.	trust path	13.	reputation
7.	Profile hijacking		

➤ Answers of Multiple Choice Questions

1.	a	2.	b	3.	c	4.	b	5.	d

❑❑❑

5 Visualization and Applications of Social Networks

Scope of the Syllabus

Graph theory - Centrality - Clustering - Node-Edge Diagrams - Matrix representation - Visualizing online social networks, Visualizing social networks with matrix-based representations - Matrix and Node-Link Diagrams - Hybrid representations - Applications - Cover networks - Community welfare - Collaboration networks - Co-Citation networks.

5.1 Graph Theory

- Graph theory is a branch of discrete mathematics. Graph is just a set of objects which are connected in some way. The objects are called vertices or nodes. Pictorially, we usually draw the vertices as circles, and draw a line between two vertices if they are connected or related. These lines are called edges or links.
- Graph theory is probably the main method in social network analysis in the early history of the social network concept. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links (for example influencers and the followers).
- Influencers on social network have been identified as users that have impact on the activities or opinion of other users by way of followership or influence on decision made by other users on the network as shown in Fig. 5.1.1.

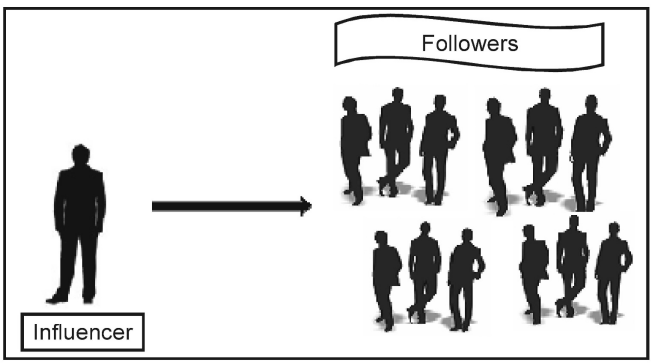
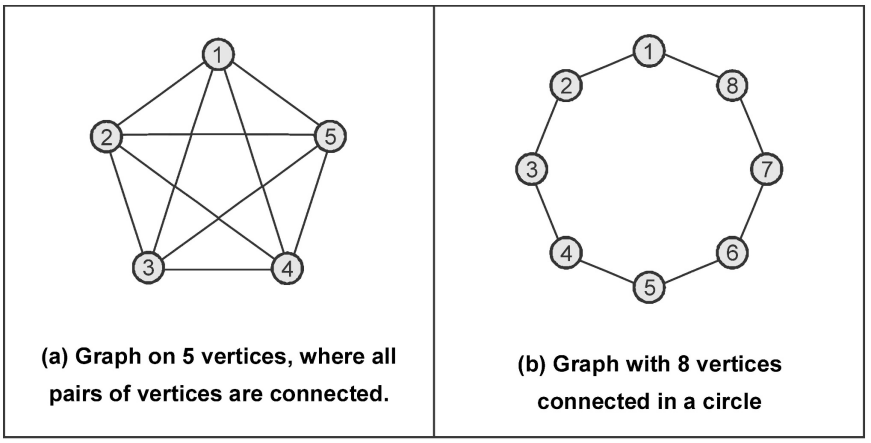


Fig. 5.1.1

- Graph theory has proved to be very effective on large-scale datasets such as social network data. This is because it is capable of bye-passing the building of an actual visual representation of the data to run directly on data matrices.



(a) Graph on 5 vertices, where all pairs of vertices are connected.

(b) Graph with 8 vertices connected in a circle

Fig. 5.1.2

- Fig. 5.1.2 shows a graph with 8 vertices connected in a circle and graph on 5 vertices, where all pairs of vertices are connected.
- Note that the physical location of the vertices in the drawings are unimportant, only which vertices are connected matters. For example, the following two graphs are the same.

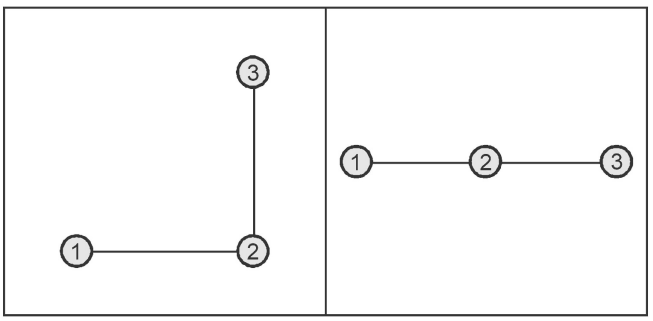


Fig. 5.1.3

- Graphs naturally arise in many ways, as they are a convenient way to visualize various situations or complex symptoms. Computer systems in a local network form a graph. So do the landlines telephone cable systems and internet routing systems.
- The World Wide Web : One can form a graph of all web pages, and make an edge from Page A to Page B if there is a hyperlink from Page A to Page B. In this case, one should consider directed edges, meaning each edge has a direction which is pictorially indicated with an arrow.
- As used in graph theory, the term graph does not refer to data charts such as line graphs or bar graphs. Instead, it refers to a set of vertices (that is, points or nodes) and of edges (or lines) that connect the vertices.

- **Graph -** A visual representation of edges and vertices.
- **Edge -** The line between two boundaries.
- **Vertex -** A point where two or more lines meet.
- Fig. 5.1.4 shows vertex and edge in graph.

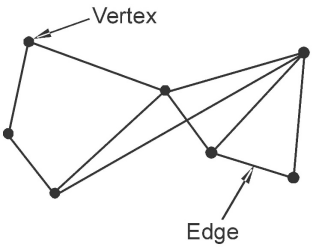


Fig. 5.1.4 : Vertex and edge in graph

- A graph G is a triple consisting of : Vertex set V(G), an edge set E(G) and a relation between an edge and a pair of vertices.

  G = (V; E)

  Where, V (or V (G)) is a set of vertices

  E (or E(G)) is a set of edges each of which is a set of two vertices (undirected), or an ordered pair of vertices (directed).
- Two vertices that are contained in an edge are adjacent; two edges that share a vertex are adjacent; an edge and a vertex contained in that edge are incident.
- **Multigraph :** Multiple edges are allowed between vertices
- **Simple graph :** A graph without loops and with atmost one edge between any two vertices
- **Pseudograph :** A graph that may contain multiple edges and graph loops
- An undirected graph is connected if every two nodes in the network are connected by some path in the network. Components of a graph (or network) are the distinct maximally connected sub-graphs.
- A directed graph is connected if the underlying undirected graph is connected. It is strongly connected if each node can reach every other node by a directed path.
- Mathematically, social networks can be represented as graphs or matrices. The nodes in a graph represent persons (or animals, organizations, cities, countries, etc) and the lines represent relationships among them. The line between persons a and b is represented mathematically like this : (a, b). The network drawn below contains these edges : (a, b), (a, e), (a, d), (b, c), and (d, c).

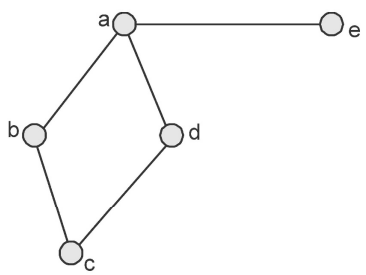


Fig. 5.1.5

- Graphs can also be valued or non-valued. A valued graph has numbers attached to the lines that indicate the strength or frequency or intensity or quantity of the tie between nodes. For example, Fig. 5.1.6 might record the amount of trade, in trillions of dollars, between some countries :

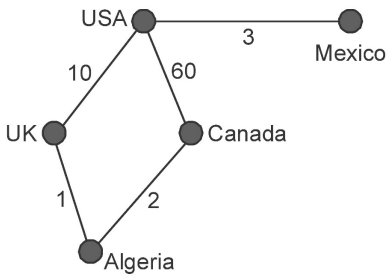


Fig. 5.1.6 : Valued graph

- If a line connects two points, they are said to be "adjacent". The two points connected by a line are called endpoints. An edge that originates or terminates at a given point is "incident" upon that point. Two edges that share a point are also said to be incident.
- A subgraph of a graph is a subset of its points together with all the lines connecting members of the subset. The subgraph of Fig. 5.1.6 that includes the UK, Canada and Algeria has two lines : (UK, Algeria) and (Algeria, Canada).
- The degree of a point is defined as the number of lines incident upon that node. In Fig. 5.1.6, the degree of USA is 3 because it has 3 ties. If a point has degree 0 it is called an isolate. If it has degree 1 it is called a pendant.
- In a directed graph, a point has both indegree and outdegree. The outdegree is the number of arcs from that point to other points.
- A node is reachable from another node if there exists a path of any length from one to the other.
- A connected component is a maximal subgraph in which all nodes are reachable from every other. Maximal means that it is the largest possible subgraph : you could not find another node anywhere in the graph such that it could be added to the subgraph and all the nodes in the subgraph would still be connected.
- For directed graphs, there strong components and weak components. A strong component is a maximal subgraph in which there is a path from every point to every point following all the arcs in the direction they are pointing. A weak component is a maximal subgraph which would be connected if we ignored the direction of the arcs.
- A cutpoint is a vertex whose removal from the graph increases the number of components. That is, it makes some points unreachable from some others. It disconnects the graph.
- The connection density in a graph is defined as the ratio of the number of edges actually present in the graph and the maximum number of edges possible.

5.2 Centrality

- Centrality is a measure indicating the importance of node in the network. Commonly, it measures the 4 P's - prestige, prominence, importance and power.

- In centrality measure was used to inspect the representation of power and influence that forms clusters and cohesiveness on social network.
- The centrality of a node in a network is a measure of the structural importance of the node. A person's centrality in a social network affects the opportunities and constraints that they face. There are three important aspects of centrality : degree, closeness, and betweenness.

➤ 1. Degree

- Degree centrality is defined as the ratio of the number of neighbours of a vertex with the total number of neighbours possible. It is simply the number of nodes that a given node is connected to. If the network consists of who knows whom, degree centrality is the number of people that a given person knows.

$$\textbf{Degree Centrality} = \frac{\textbf{Degree of the vertex(k)}}{\textbf{Total number of nodes in the network (N) – 1}}$$

- The variance of the distribution of degree centrality in a network gives us the centralization of the network. One can see that a star network is an ideal centralized network, whereas a line network is less centralized.

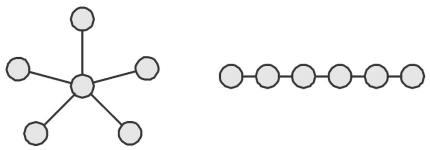


Fig. 5.2.1 : Star network and line network

- In general, the greater a person's degree, the more potential influence they have on the network, and vice-versa. For example, in a gossip network, a person who has more connections can spread information more quickly, and will also be more likely to hear more stuff. This can be both good and bad.
- The greater a person's degree, the greater the chance that they will catch whatever is flowing through the network, whether good or bad.

➤ 2. Closeness

- Closeness centrality is defined as the total graph-theoretic distance to all other nodes in the network. For example, in Fig. 5.2.2 node “e” has a closeness score of 8 because it is one link away from “a”, two links away from “b” and “d”, and three links away from “c”. The bigger the number the LESS central because they are farther away from everyone.

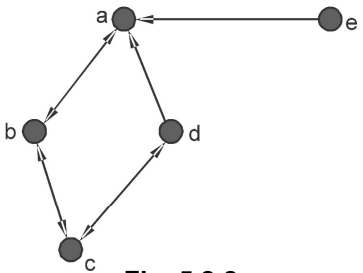


Fig. 5.2.2

- When a node has a low closeness score, it tends to receive anything flowing through the network very quickly. This is because the speed with which something spreads in a network is a function of the number of links in the paths traversed.
- Since nodes with low closeness scores are close to all nodes, they receive things quickly. Once again, whether this is good or bad depends on the situation. In the case of information about what's happening in the company, this is usually good. In the case of a new disease that is spreading, it is very bad to be one of the first people to get it .

➤ 3. Betweenness

- The degree of a node is not the only measure of the importance of a node in the network, and this centrality measure addresses this fact. This concept was introduced by Linton Freeman.
- Vertices that have a high probability of occuring on a randomly chosen shortest path between two nodes are said to have high betweenness centrality.
- Model based on communication flow : A person who lies on communication paths can control communication flow, and is thus important. Betweenness centrality counts the number of shortest paths between i and k that actor j resides on.
- This measures the number of times a given vertex u lies on a (shortest length) path between other vertices v1 and v2 , and gives a higher score the more times u appears.

➤ Katz centrality

- Katz centrality can be used to compute centrality in directed networks such as citation networks and the World Wide Web. Katz centrality is more suitable in the analysis of directed acyclic graphs where traditionally used measures like Eigenvector centrality are rendered useless.
- Katz centrality can also be used in estimating the relative status or influence of actors in a social network. Each node is provided a small amount of centrality irrespective of its position.

➤ Interpretation of measures

Centrality measure	Interpretation in social networks
Degree	How many people can this person reach directly ?
Betweenness	How likely is this person to be the most direct route between two people in the network ?
Closeness	How fast can this person reach everyone in the network ?
Eigenvector	How well is this person connected to other well-connected people ?

➤ **Centrality and power**

Power aspect name	Definition	Influences
Degree	Number of ties for an actor	Having more oppurtunities and alternatives
Closeness	Length of paths to other actors	Direct bargaining and exactors change with other actors
Betweenness	Lying between each other pairs of actors	Brokering contacts among actors to isolate them or prevent connections.

5.2.1 Page Rank

- PageRank : A method for rating the importance of web pages objectively and mechanically using the link structure of the web.
- PageRank was developed by Larry Page and Sergey Brin. It is first as part of a research project about a new kind of search engine. That project started in 1995 and led to a functional prototype in 1998.
- The Page Rank algorithm, used in the Google search engine considers that users have an absolute preference among Web pages : it assumes that the more a Web page is visited, the more it is appreciated by the users.
- To measure the popularity of the pages, it is not possible to have access to the logs of the servers, but a reasonable assumption is that the preference of users is reflected in the hypertext structure: a link toward a Web page is often an indication that this page is acknowledged by someone as a good source of information.
- A simple way to implement this idea would be to count the number of times a Web page is cited.
- The idea behind PageRank is that a user who crawls the Web by selecting the hyperlinks at random is more likely to visit certain Web pages than others, simply because there are more possible ways by which the user can reach these pages.
- It is possible to model the behavior of a "random" surfer as a Markov process, where the states of the system are each of the Web pages. The measure of popularity of a Web Page, its PageRank, is given by the stationary probabilities of this Markov process - the limit probability that the user will be on a certain page.
- Searching with PageRank : Two search engines :
  a. Title-based search engine
  b. Full text search engine

➤ **Title-based search engine**

- It searches only the "Titles". Finds all the web pages whose titles contain all the query words.

- Sorts the results by PageRank.
- Very simple and cheap to implement
- Title match ensures high precision and PageRank ensures high quality

➤ **Full text search engine**

- Also called Google. It examines all the words in every stored document and also performs PageRank.
- More precise but more complicated.

5.3 Clustering

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Let us consider following Fig. 5.3.1.

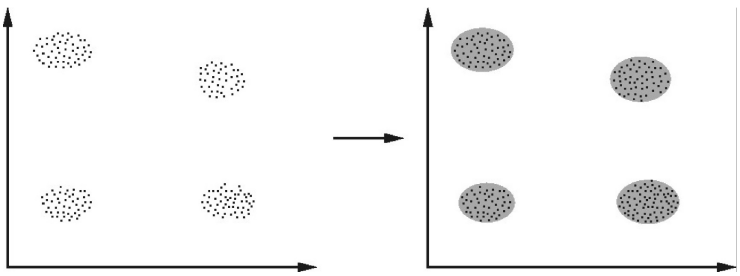


Fig. 5.3.1

- In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering**.
- **Conceptual clustering :** Two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.
- Clustering methods are usually categorized according to the type of cluster they produce. The clustering methods are categorized as :
  1. **Hierarchical methods :** These types cluster produces the output list of cluster. Small clusters of highly similar documents nested within larger clusters of less similar documents.
  2. **Non-hierarchical methods :** This method produced unordered lists.
- Other clustering methods are **exclusive cluster** and **overlapping cluster**. In the first case (exclusive cluster) data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. The **overlapping clustering** uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

- Four of the most used clustering algorithms :
  a) K-means b) Fuzzy C-means
  c) Hierarchical clustering d) Mixture of Gaussians
- Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly mixture of Gaussian is a probabilistic clustering algorithm.

5.4 Node-Edge Diagrams

- Visualization plays a crucial role of linking the human vision and computer, helping identify patterns, and extracting insights from large amounts of information.
- Graphs are generally visualized as node-link diagrams, in which dots depict the nodes, joined by lines or curves for the edges
- Graphs depicted as node-link diagrams are widely used to show relationships between entities. However, node-link diagrams comprised of a large number of nodes and edges often suffer from visual clutter
- For visualizing social networks, some visual representations are considered appropriate to present network structures, such as node-edge diagrams and matrix representations.
- A node-edge diagram is used to visualize social networks. Using this diagram, user can perform many network analysis tasks, such as component size calculation, centrality analysis, and pattern sketching.
- Many node-edge layouts have been presented to place the nodes in the graph for users to clearly recognize the structure of the social network. Different layouts have their own merits and demerits to display the network graph depending on the size, complexity, and structure of the social network.
- For example, some layouts are suitable to display social networks in a moderate size, but they are not suitable for showing larger networks. Three kinds of layouts, namely, random layout, force-directed layout, and tree layout, are described to explain the node-edge diagrams.

➤ 1. Random layout

- A random layout is used for putting the nodes at random geometric locations in the graph. It cannot give clear visualization results when number of nodes immensely increases.
- Random layout algorithm can efficiently draw the social network graph in linear time, i.e. O (N). Fig. 5.4.1 shows random layout graph.
- Random graphs have been proposed as a possible model to take into account the structural characteristics of instances that appear in many practical applications

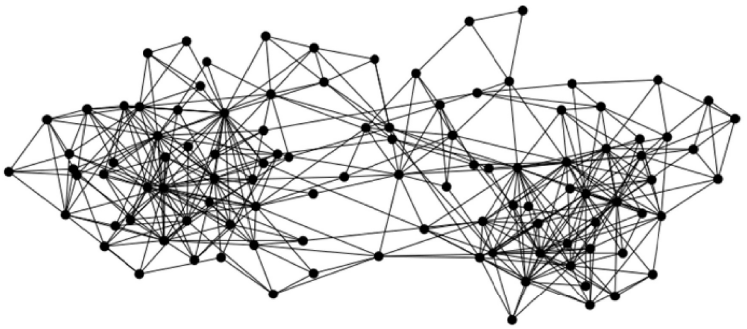


Fig. 5.4.1 : Random layout graph

➤ 2. Force - directed layout

- Force-directed layout is also known as a spring layout, which simulates the graph as a virtual physical system. We imagine the nodes as physical particles that are initialized with random positions, but are gradually displaced under the effect of various forces, until they arrive at a final position.
- Spring layout is based on a cost (energy) function, which maps different layouts of the same graph to different non-negative numbers.
- The forces are defined by the chosen algorithm, and typically seek to position adjacent nodes near each other, but not too near.
- This layout may take hundreds of iterations to obtain a stable layout, the running time is atleast O(N log N) or O(E) where N is the number of nodes and E is the number of edges.
- Compared with a random layout, the running cost of a force-directed layout is much higher than that of a random layout, especially when the number of nodes is large.
- It is not suitable for graphs larger than hundreds of nodes. Large graphs often make the energy function very hard to reach minimum.
- Force-directed algorithms show a lack of predictability, which means two different runs of the same algorithm with a same input graph may be unlike one another

➤ 3. Tree layout

- Basic tree layout is to choose a node as the root of tree, and the nodes connected to the root become children of the root node. Tree layout is more tractable and easier to understand.
- A tree layout can display a more structural layout than graph layouts. Fig 5.4.2 shows tree layout. Node-link layouts use links between nodes to indicate the parent-child relationships.
- Generally, radial views, including its variations, share a common characteristic : the focus node is always placed at the center of the layout, and the other nodes radiate outward on separated circles.

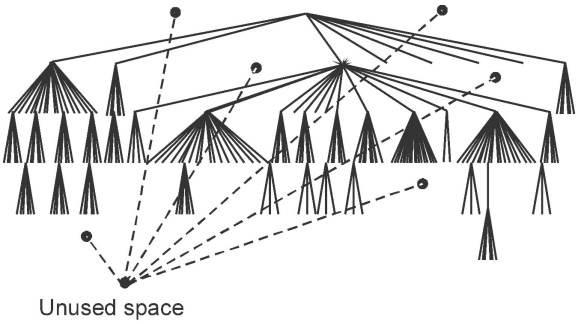


Fig. 5.4.2 Tree layout

- Balloon layout is similar to radial layout. Balloon layouts are formed where siblings of sub-trees are placed in circles around their father node. This can be obtained by projecting cone tree onto the plane.

5.5 Matrix Representation

- The last solution to visualize large diagrams is to resort to a different representation than node-link diagrams. An obvious choice is to use the adjacency matrix representation.
- Graphs can be presented by their connectivity matrixes. Each row and each column corresponds to a node.
- Edge attributes are encoded as visual characteristics of the glyphs. Such as color, shape, and size.
- If vertex A is connected to vertex B, the cell at the intersection of the line of A and the column of B is marked. Since vertices are represented both in rows and columns, there are two cells corresponding to a pair of vertices, making it possible to represent directed edges.
- Traditionally, a numerical value marks the connection (0 if no connection, 1 if there is one, n if the edge is weighted).
- Replacing numerical values by visual indicators and reordering rows and columns dramatically improves the readability of tables and matrices.
- An adjacency matrix contains one row and one column for each node of a network. Given two nodes i and j, the cells located at (i, j) and (j, i) in the matrix contain information about the edge(s) between the two nodes. Typically, each cell contains a boolean value indicating if an edge exists between the two nodes.
- If the graph is undirected, the matrix is symmetric, i.e., the two cells (i, j) and (j, i) correspond to the same edge. If the graph is directed, however, the matrix is not symmetric.
- Visualizing a network as a matrix has the advantage of eliminating all edge crossings, since the edges correspond to non-overlapping cells.
- Matrices have the added advantage of also being able to display information related to each edge within the cells of the matrix. For example, if the edges are weighted, this weight can be shown in the color of the cell. Cells can also contain small graphics or glyphs.

- Disadvantage of using adjacency matrices, is that the space they require is $O(N^2)$ where N is the number of nodes.
- Both the matrix and node-link representations support the analysis of the network at different levels of details. if an analyst is looking for an overview of the network to identify its main communities, the matrix is the best option. When a more detailed analysis is required, to identify actors bridging two communities for example, node-link diagrams constitute a better alternative.
- The matrix and node-link representations are synchronized to combine their advantages and ease the identification of visual patterns. Selecting a row or column in the matrix highlights the corresponding node in the other representation.

5.6 Visualizing Online Social Networks

- The advent of the internet has given rise to many forms of online sociality, including e-mail, usenet, instant messaging, blogging, and online dating services. In 2003, another form of online community acquired stunning popularity : online social networking services.
- Versatile visualization skills are employed to facilitate analyzing online social networks. visualizations of online social networks were developed according to the attributes of network sociality to present their network structure
- Visualization of social networks has a rich history, particularly within the social sciences, where node-link depictions of social relations have been employed as an analytical tool since atleast the 1930s.
- Networks can be arranged on a map to represent the geographic distribution of a population. color, size, and shape have been used to encode both topological and non-topological properties such as centrality, categorization, and gender.
- There are two obvious criteria for the quality of social network visualizations :
  1. The information manifest in the network represented accurately ?
  2. Is this information conveyed efficiently ?
- The following three aspects should be carefully thought through when creating network visualizations :
  1. The substantive aspect the viewer is interested in,
  2. The design
  3. The algorithm employed to realize the design (artifacts, efficiency, etc.)

5.6.1 Web Communities

- Web community is a web site where specific content or links are only available to its members. A web community may take the form of a social network service, an internet forum, a group of blogs, or another kind of social software web application.
- Web community is a collection of web pages in which each member page has more hyperlinks within the community than outside the community.

- Semantic web platform makes knowledge sharing and re-use possible over different applications and community edges. Discovering the evolvement of Semantic Web (SW) enhances the knowledge of the prominence of Semantic Web Community and envisages the synthesis of the Semantic Web.
- Community membership is a function of both a Web page's outbound hyperlinks and all other hyperlinks on the Web because the rest of the Web collectively forms a page's inbound hyperlinks.
- The visualization techniques are mainly introduced to deal with the complex social relationships based on human-centric or user-centric views. To build a visualization system that end-user of social networking services could use to facilitate discovery and increased awareness of their online community.
- As the development of Semantic Web, a project called FOAF (Friend-of-a-Friend) was proposed to visualize such human-centric social relationships based on Semantic Web social metadata. With XML/RDF format, the FOAF relations can be explicitly defined for further social network analysis and visualization.
- To predict whether one person is a friend of another, we rank all users by their similarity to that person.
- Similarity is measured by analyzing text, links, and mailing list. If we are trying to evaluate the likelihood that user A is linked to user B, we sum the number of items the two users have in common.
- Items that are unique to a few users are weighted more than commonly occurring items. The weighting scheme we use is the inverse log frequency of their occurrence.
- The Semantic Web and social network models support one another. On one hand, the Semantic Web enables online and explicitly represented social information; on the other hand, social networks, especially trust networks, provide a new paradigm for knowledge management in which users “outsource” knowledge and beliefs via their social networks
- The SixDegrees.com website was an early representative created on the basis of the Web interaction model during 1997 and 2001. Since the start of SixDegrees.com, various social network websites and Web-based dating services have been established to provide people more convenient ways to build up their social relationships and communities.
- In addition, many social network websites are developed with interactive visualization interfaces to facilitate people connecting their communities and maintaining social relationships.
- **Friendster** was designed to be an online dating site, complete with profiles, demographic and interest driven search, and a private messaging system. What made Friendster unique was its articulated social networking component and testimonial feature.
- Users were asked to declare “friends” on the system whose pictures would also appear on the profile when the friends confirmed the relationship. Friends could write testimonials that would also appear on the profile.

- Both the friends and testimonials were intended to signal additional information about the person's character for those interested in dating the person.
- Yet, when the early adopters began to use the service, they did not view it as a dating service, but a site where they could gather and communicate with their friends, surf for entertaining profiles and explore public displays of identity and relationships.
- To build a visualization system that end-user of social networking services could use to facilitate discovery and increased awareness of their online community.

5.6.2 Email Group

- Email service is one of the most popular applications that people often use to connect each other and deliver messages. For analyzing the social structures of the daily email activities, visualization techniques are employed to explore different patterns.
- Soylent as well as Post History and Social Network Fragments are visualizations designed to reveal social patterns in user activity (primarily based on email), with substantial focus on temporal aspects. While these tools help users manage their own contact lists, they do not expose them to individuals beyond their egocentric network.
- Soylent visualizations are based on email histories. Messages are read from a mail server, and saved to a database. The database is then analyzed and downloaded in order to be visualized.
- Personal social information is a private and important thing. The Soylent system is designed to provide additional information to its users without revealing anything to outside parties.
- In order to do that, the system reads mail off of the user's personal account, and saves it to a private database, located on the user's own machine. No one can see this information in any form except the user whose mail is being viewed.

5.7 Visualizing Social Networks with Matrix-Based Representations

- The node-link diagrams are more effective for very small (under 20 vertices) and sparse networks whereas matrices outperform them otherwise except when the task is to follow paths in the network.

➤ Matrix Advantages

1. It provides powerful overview visualization. It takes less time and always readable.
2. This visualization does not suffer from node overlapping.
3. Matrices do not suffer from link crossing each other.
4. Matrices show all possible pairs of vertices, they can highlight the lack of connections and also the directedness of the connections.

5.7.1 Matrix and Node-Link Diagrams

- MatrixExplorer is based on two representations : Matrix-based and node-link diagram. Node-link and matrix visualizations are synchronized in order to let the user work with both representations; our goal is to allow them to switch smoothly from one to the other.

- Fig. 5.7.1 shows visual patterns in Matrix and Node-link representations of social networks.

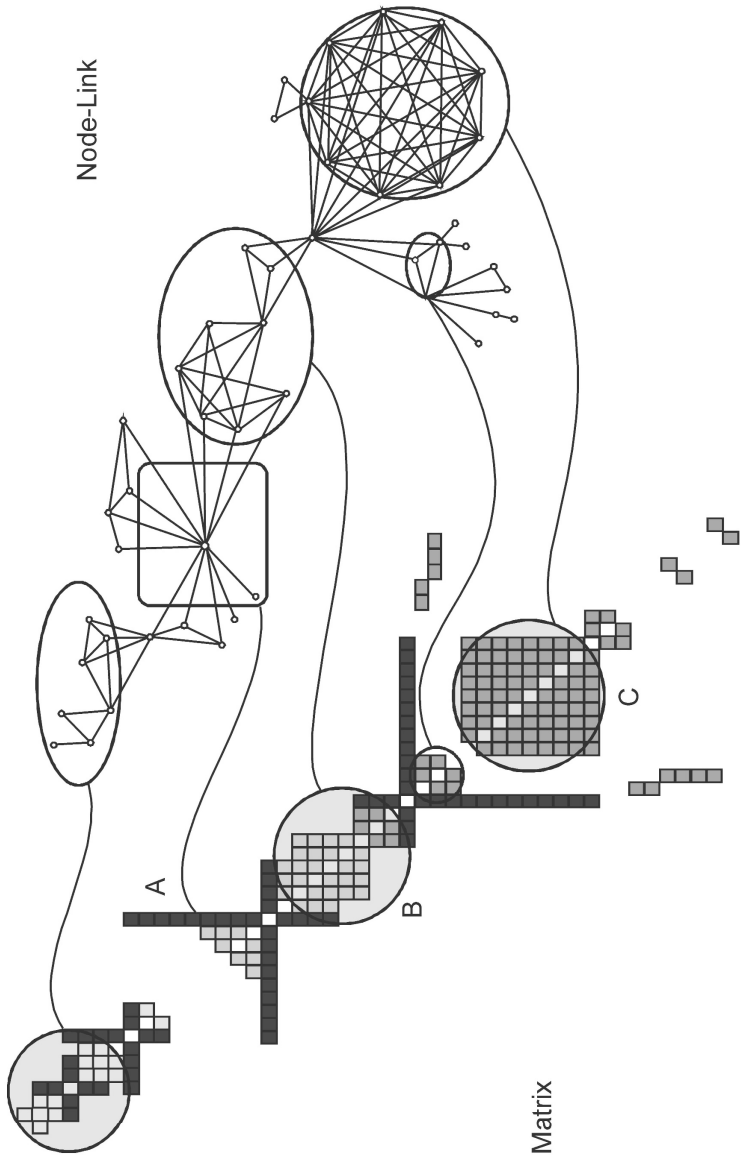


Fig. 5.7.1 : Visual patterns in matrix and node-link representations of social networks

- Multiple visualizations are synchronized by selection and filtering. If a user selects a set of actors in the matrix, this same set will be selected in all other visualizations (selection) and data filtered in one visualization will disappear from all others (filtering).
- Selection improves the transition from one representation to the other and constitutes the core of the coupling. Filtering preserves the coherence of the visualizations by presenting the same data, even if the attributes visualized are different.

- The process to obtain both representations follows :
  1. The user first automatically ordered the matrix, identified clusters (communities) and attributed colors to identify them.
  2. User then switched to a node-link diagram, displaying the community colors and laying the network out manually in order to better visualize how communities are linked and organized.
  3. Finally, moving back and forth between both representations, he identified the global structure of the network.
- To interactively manipulate matrix and node-link representations, the following set of tools are used :
  1. Interactive specification of visual attributes. The user controls the mapping data-visual encoding by entering values in a text field or selecting a value in a list. Visual attributes of nodes, rows or columns such as label, color, transparency or size as well as attributes of links or cells such as thickness, color or texture may be associated to a data attribute.
  2. Interactive layout and reordering. Users may directly move a node or a row/column in both representations to change its position or order.
  3. Automatic layout and reordering techniques.
  4. Computer-assisted layout and reordering techniques.
  5. Interactive filtering. This functionality allows filtering actors or connections ac-cording to a selection or by selecting a specific value of a data attribute from a list.
  6. Interactive clustering.

5.7.2 Hybrid Representations

- List of drawback
  1. It requires a large amount of display space
  2. Switching from one representation to the other may induce high cognitive load to the user
- Finding the shortest path between two given actors is easy using node-link diagram. Here user can quickly find the multiple paths and select the shortest path. These tasks being very common in social network analysis. So hybrid representation is used to solve the problem in matrices.
- MatLink displays the full graph using a linearized node-link representation and called as the full linear graph. Its links are curved lines drawn interior to the vertex displays at the top and left edges of the matrix.
- Links are drawn over the matrix cells, using transparency to avoid hiding them. Longer links are drawn above shorter ones.
- The linear graph conveys detailed and long-range structure together without hiding any detail of the matrix : a feeling for link densities and sub-graphs, but also paths and cut points

➤ Merging Matrix and Node-Link Diagram

- The need to visualize large social networks is growing as hardware capabilities make analyzing large networks feasible and many new data sets become available. Unfortunately, the visualizations in existing systems do not satisfactorily answer the basic dilemma of being readable both for the global structure of the network and also for detailed analysis of local communities.
- To address this problem, NodeTrix is used. Hybrid representation for networks that combines the advantages of two traditional representations : Node-link diagrams are used to show the global structure of a network, while arbitrary portions of the network can be shown as adjacency matrices to better support the analysis of communities.
- A key contribution is a set of interaction techniques. These allow analysts to create a NodeTrix visualization by dragging selections from either a nodelink or a matrix, flexibly manipulate the NodeTrix representation to explore the dataset, and create meaningful summary visualizations of their findings.
- NodeTrix is a hybrid representation of networks based on the node-link diagram where communities can be represented as matrices. Intra-community relationships use the adjacency matrix representation while inter-community relationships use normal links.
- NodeTrix is built on the InfoVis Toolkit and uses its rendering mechanism to create the visualization. The rendering mechanism involves a pipeline of renderers which makes it simple to draw a matrix over a standard node.
- To display links in NodeTrix, three options are used : Displaying only aggregated links, displaying only the underlying links, or displaying both.
- Displaying aggregated links provides simple visual feedback on how communities interact. Moreover, an aggregated attribute can be mapped to a visual variable (e.g. color, thickness, opacity) of this edge.
- Displaying each underlying edge provides connectivity details and enables visualization of the attributes of each edge independently, but at the cost of many more links and potential crossings.
- Drawback of NodeTrix is the concrete representation of communities, making it impossible to place an actor in two different communities.
- NodeTrix can be used both for exploring and presenting publications data.

5.8 Applications

- 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 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. Management consultants use this methodology with their business clients and call it Organizational Network Analysis [ONA]. ONA allows you to x-ray your organization and reveal the managerial nervous system that connects everything.

1. Organizational Issues : Organizations can be viewed as social groupings with relatively stable patterns of interaction over time. The social network approach views, organizations in society as a system of objects (e.g. people, groups, organizations) joined by a variety of relationships. Not all pairs of objects are directly joined, and some are joined by multiple relationships. Network analysis is concerned with the structure and patterning of these relationships and seeks to identify both their causes and consequences.
2. The SNA in e-learning recommendation systems is also proved to be useful to present learners with the proper documentation choice without having sufficient personal experience or knowledge of available informatics.

5.8.1 Covert Networks

- The covert networks are hidden, the actors of such network does not disclose their information to the external world.
- Increased interest in studying 'covert networks' using SNA, but many unresolved substantive and methodological questions : -
  1. Not least questions about the concept itself : 'secrecy' as the defining feature, but secrecy of what ? From whom ? When ? -
  2. What makes a network 'covert', and how covertness is achieved, varies and demands sociological attention
  3. Terrorist networks and criminal networks are prime examples, but many others.
- Aim to create an archive of covert network datasets and then use it for theoretical exploration, empirical analysis, and methodological development.
- A covert network is a social network which has one or many elements of secrecy about it.
- Network members may try and keep their identities secret; the network may form around activities which have to be kept secret because they are illegal or dangerous , or for other reasons.
- An important area in SNA is the key player detection. Key player is defined as the most important node in a social network. Centrality is a key theory in the study of social networks in order to study organizational and team behavior. Central individuals control information flow and decision making within a network .
- Along with key player detection, outlier detection is also important to predict any abnormal activity. Outlier detection deals with detection of patterns from data which do not match expected normal behavior. These anomalous patterns are often known as outliers, anomalies, discordant observations, and so forth in different application domains.
- Outlier detection is a well-researched area having an immense use in a wide range of applications like fraud detection, insurance, intrusion detection in cyber security, fault detection in security critical systems, military surveillance for enemy activities, and so on.
- Terrorist groups and organizations are hidden networks which does not disclose their identity, generally the data to build and complete such networks is gathered from publicly available resources such as news papers.

- Now a days Web resources such as blogs, emails etc. are also used for hidden communication. Hence, various data mining and social network analysis techniques are employed to extract necessary information to detect terror.

5.8.2 Community Welfare

- Social network can be established through family ties, friendship, common interest, financial exchange, dislike, sexual relationship, beliefs, knowledge or prestige. Social network is mostly complex operating on different levels.
- It determines one's social contacts and the value of one's social capital. It can be developed through internet sources such as facebook, hyves, skype, linked in; it can also be developed through business network, sports activities, church associations, funeral gatherings, social clubs and pubs among others.
- Fig. 5.8.1 shows social network community structure.

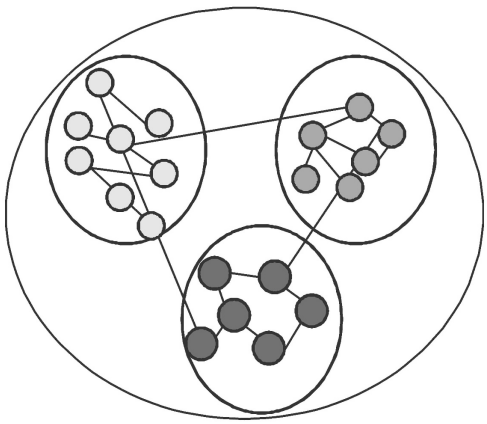


Fig. 5.8.1 : Social network community structure

- Among the migrant community in the Netherlands, there exist different platforms and network of inter relationships; where migrants are engaged with one another to share ideas and broaden their horizon.
- For instance, among the Ghanaian community within the Hague, there exists the Asanteman club which seeks to involve all Ashanti's in Netherlands.
- Recogin is a platform organization operating within a social network for Ghanaian migrant organizations in Netherlands, although its office is based in Amsterdam.
- Mass surveillance is one of the modern practices undertaken by some organizations and governments to monitor the behavior of suspected people of population

5.8.3 Collaboration Networks

- A collaborative network is a network consisting of a variety of entities (e.g. organizations and people) that are largely autonomous, geographically distributed, and heterogeneous in terms of their operating environment, culture, social capital and goals, but that collaborate to better achieve common or compatible goals, and whose interactions are supported by computer networks.

- Coauthor ship is more common in the natural sciences than in the social sciences, but has been increasing steadily across all fields.
- Co-authorship creates a social network, the study of which allows us to understand the structure of scientific collaborations, some of the characteristics of a particular discipline and to identify the invisible colleagues and social groups that exist in all scientific fields and status of individual researchers
- Several explanations have been given for the increase in coauthorship over time. The examples of co-authorship networks are Wikipedia article authors, network of the pacific Asia Conference on Information Systems, network of European Conference on Information Systems (ECIS) etc.
- SNA on these networks has been conducted to understand the research community which produces the research knowledge
- Funding requirements, particularly in large lab settings, might induce collaboration. While social scientists are rarely as dependent on labs, the rise of large-scale data collection efforts suggests a similar team-production model.
- Training differences between disciplines might also account for coauthorship differences. Advanced work by PhD students in the natural sciences is usually closely related to an advisor's work, and commonly results in collaboration. Social science students, in contrast, tend to work on projects that are more independent.

5.8.4 Co-Citation Networks

- Co-citation is used as a measure of similarity between two objects. Co-citation analysis helps to understand the status and structure of scientific research.
- Co-citation analysis result shows that authors who belong to cliques are the most cited, while the page-rank scores emphasized the importance of who cites an article as articles cited by popular authors had higher page-rank scores and the most cited articles were not necessarily the most important depending on who cites them.
- Co-citation analysis is an example of a deconstruction assignment method that uses the references at the end of the document.
  1. References in a document are identified.
  2. The relatedness between these references are calculated.
  3. The references are clustered using a transform of the co-occurrence matrix. And finally, the original documents are assigned to these reference clusters.
- Co-citation analysis studies structures of scientific research, based upon citations and co-citations. It enables researchers identify groups of scientists and their publications and to draw conclusions about the inner structure of research disciplines , schools and paradigms.
- Co-citation occurs when more than one reference or author appear in the same bibliography. It is a measure of the similarity of content of the references or authors. The proximity of any two publications in terms of content is determined by the number of co-citations.

- Co-citation analysis is a form of biblio-metrics or quantitative bibliography which generally involves counting citations to other publications in a body of literature and developing statistical distributions with these counts.
- Citation counts only give an idea of "who cites whom" but can't identify networks of interconnections amongst scholars. This is where citation analysis comes in. It is a document coupling technique which measures the number of documents that have cited any given pair of documents.

5.9 Questions with Answer

5.9.1 Two Marks Questions with Answer

Q. 1 What is graph theory ?

Ans. : Graph is formed by vertices and edges connecting the vertices. Graph theory is a pair of sets (V, E), where V is the set of vertices and E is the set of edges, formed by pairs of vertices.

Q. 2 What is centrality ?

Ans. : In centrality measure was used to inspect the representation of power and influence that forms clusters and cohesiveness on social network.

Q. 3 Define degree centrality.

Ans. : Degree centrality is defined as the ratio of the number of neighbours of a vertex with the total number of neighbours possible.

Q. 4 What is closeness ?

Ans. : Closeness centrality is defined as the total graph-theoretic distance to all other nodes in the network.

Q. 5 What is convert network ?

Ans. : A covert network is a social network which has one or many elements of secrecy about it.

Q. 6 Define collaboration network.

Ans. : Collaboration network consists groups of persons working together to perform particular activity and studying human collaboration is an important topic in sociology.

Q. 7 What is co-citation ?

Ans. : Co-citation is used as a measure of similarity between two objects. Co-citation analysis helps to understand the status and structure of scientific research.

Q. 8 What is use of node-edge diagram ?

Ans. : A node-edge diagram is used to visualize social networks. Using this diagram, user can perform many network analysis tasks, such as component size calculation, centrality analysis, and pattern sketching.

Q. 9 What is web community ?

Ans. : Web community is a collection of Web pages in which each member page has more hyperlinks within the community than outside the community

5.9.2 Fill in the Blanks

Q. 1 Graph is a visual representation of edges and ----------.

Q. 2 ---------- centrality is defined as the ratio of the number of neighbours of a vertex with the total number of neighbours possible.

Q. 3 ---------- distance is defined as the least number of connections that must be traversed to get between any two nodes

Q. 4 The random surfer visits a web page with a certain probability which derives from the page's ---------.

Q. 5 Co-citation is used as a measure of --------- between two objects.

Q. 6 ---------- is a hybrid representation of networks based on the node-link diagram where communities can be represented as matrices.

Q. 7 Degree centrality is the sum of all other actors who are directly connected to --------.

Q. 8 A ------- in a graph is a sub-graph in which any node is directly connected to any other node of the sub-graph

Q. 9 The covert networks are ---------, the actors of such network does not disclose their information to the external world.

Q. 10 NodeTrix is a hybrid representation of networks based on the ---------- diagram where communities can be represented as matrices.

Q. 11 Force-directed layout is also known as a -------- layout, which simulates the graph as a virtual physical system.

Q. 12 ---------- centrality is defined as the total graph-theoretic distance to all other nodes in the network.

Q. 13 A random layout cannot give clear visualization results when number of nodes immensely ----------.

Q. 14 The process of grouping a set of physical or abstract objects into classes of similar objects is called ----------.

5.9.3 Multiple Choice Questions

Q. 1 Which of the following NOT centrality measures?

(a) Degree (b) Graph (c) Katz (d) closeness

Q. 2 Force-directed layout is also known as a ---------- layout, which simulates the graph as a virtual physical system.

(a) Tree (b) force-directed

(c) Spring (d) None of these

Q. 3 Computing betweenness Centrality of a given node involves computing which of the following?:

(a) The number of shortest paths between the given node and the highest degree node.

(b) The number of longest paths between the given node and the highest degree node.

(c) The number of shortest paths that pass through the given node.

(d) The number of longest paths that pass through the given node.

Q. 4 The most important centrality measures are :

(a) degree centrality (b) closeness centrality

(c) between-ness centrality (d) All of these

Q. 5 If the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a ----------.

(a) dendrogram (b) nearest-neighbor clustering algorithm

(c) minimal spanning tree algorithm (d) single-linkage algorithm

Q. 6 Node-edge diagram is represented by ----------

(a) Random layout (b) Tree layout

(c) Web layout (d) (a) and (b)

➤ Answers of Fill in the Blanks

1.	vertices	8.	clique
2.	Degree	9.	hidden
3.	Geodesic	10.	node-link
4.	pagerank	11.	spring
5.	similarity	12.	Closeness
6.	Node Trix	13.	increases
7.	ego	14.	clustering

➤ Answers of Multiple Choice Questions

1.	b	2.	c	3.	c	4.	d	5.	d	6.	d

❑❑❑

Notes

SOLVED MODEL QUESTION PAPER

(As Per New Syllabus)

Social Network Analysis

Semester – VIII (CSE / IT) Professional Elective-IV

Time : Three Hours] **[Maximum Marks : 100**

Instructions :

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**

Part A - (10 × 2 = 20 Marks)

Q.1 *What is a social network?* **[Refer Two Marks Q.4 of Chapter 1]** **[2]**

Q.2 *What is two mode network?* **[Refer Two Marks Q.9 of Chapter - 1]** **[2]**

Q.3 *What is Direct Binary Relations?* **[Refer Two Marks Q.16 Chapter - 2]** **[2]**

Q.4 *Define ontology.* **[Refer Two Marks Q.2 Chapter - 2]** **[2]**

Q.5 *Explain dynamic social network.* **[Refer Two Marks Q.10 Chapter - 3]** **[2]**

Q.6 *What is dendrogram?* **[Refer Two Marks Q.3 Chapter - 3]** **[2]**

Q.7 *What is reality mining?* **[Refer Two Marks Q.1 Chapter - 4]** **[2]**

Q.8 *What is sybil attacks?* **[Refer Two Marks Q.4 Chapter - 4]** **[2]**

Q.9 *What is convert network?* **[Refer Two Marks Q.5 Chapter - 5]** **[2]**

Q.10 *What is web community?* **[Refer Two Marks Q.9 Chapter - 5]** **[2]**

Part B - (5 × 13 = 65 Marks)

Q.11 a) **i)** *What is semantic web? What are the limitation of current web? Explain benefits of the Semantic Web.* **[Refer section 1.1]** **[8]**

ii) *What is Social Network Analysis? Explain Principles of Social Network Analysis.* **[Refer section 1.3]** **[5]**

OR

b) **i)** *What is Blogs and Online Communities? Explain in details.* **[Refer section 1.5.2]** **[8]**

ii) *Draw and explain Generic Architecture of Semantic Web Applications .* **[Refer section 1.6.1]** **[5]**

Q.12 a) *What is an Ontology? Explain its role in the Semantic Web. Why Develop Ontology ?* **[Refer section 2.2]** **[13]**

OR

(b) *Describe various methods of Aggregating and Reasoning with Social Network Data.* **[Refer section 2.6]** **[13]**

Q.13 a) **i)** *How to detect communities in social networks?* **[Refer section 3.2]** **[8]**

ii) *Explain Decentralized Online Social Networks.* **[Refer section 3.8]** **[5]**

OR

(b) *Explain the following methods of detecting communities :*

1. Divisive algorithms 2. Modularity optimization

3. Spectral algorithms **[Refer section 3.5]** **[13]**

Q.14 a) **i)** *What is do you mean trust network analysis? Explain* **[Refer section 4.4]** **[8]**

ii) *Comments on "Trust in Online Environment "* **[Refer section 4.3.1]** **[5]**

OR

(b) *Explain various methods of enabling new human experiences.* **[Refer section 4.2]** **[13]**

Q.15 a) **i)** *What is graph theory? Explain* **[Refer section 5.1]** **[8]**

ii) *Write short note on Covert Networks.* **[Refer section 5.8.1]** **[5]**

OR

b) *Explain Visualizing Social Networks with Matrix-Based Representations.* **[Refer section 5.7]** **[13]**

Part C - (1 × 15 = 15 Marks)

Q.16 a) *Explain Node-Edge Diagrams of social network.* **[Refer section 5.4]** **[15]**

OR

b) **i)** *Explain Resource Description Framework (RDF) and RDF Schema* **[Refer section 2.3.1]** **[10]**

ii) *Comments on Friend of a friend* **[Refer section 3.8]** **[5]**

❑❑❑