SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT – I – SOCIAL NETWORK ANALYSIS – SITA3005

UNIT I

INTRODUCTION

Introduction to Semantic Web: Limitations of current Web - Development of Semantic Web - Emergence of the Social Web Social Network Analysis: Social Networks Perspective - Analysis of Network Data - Interpertation of Network Data - Social Network Analysis in the Social and Behavioral Sciences - Metrics in social network analysis

Semantic Web

- The Semantic Web is the application of advanced knowledge technologies to the Web and distributed systems in general.
- Information that is missing or hard to access for our machines can be made accessible using *ontologies*.
- Ontologies are formal, which allows a computer to emulate human ways of reasoning with knowledge.
- Ontologies carry a social commitment toward using a set of concepts and relationships in an agreed way.
- The Semantic Web adds another layer on the Web architecture that requires agreements to ensure interoperability.

LIMITATIONS OF THE CURRENT WEB

 The current Web has its limitations when it comes to:

- finding relevant information
- extracting relevant information
- combining and reusing information

 There is a unusual ability to adapt to the limitations of our information systems.

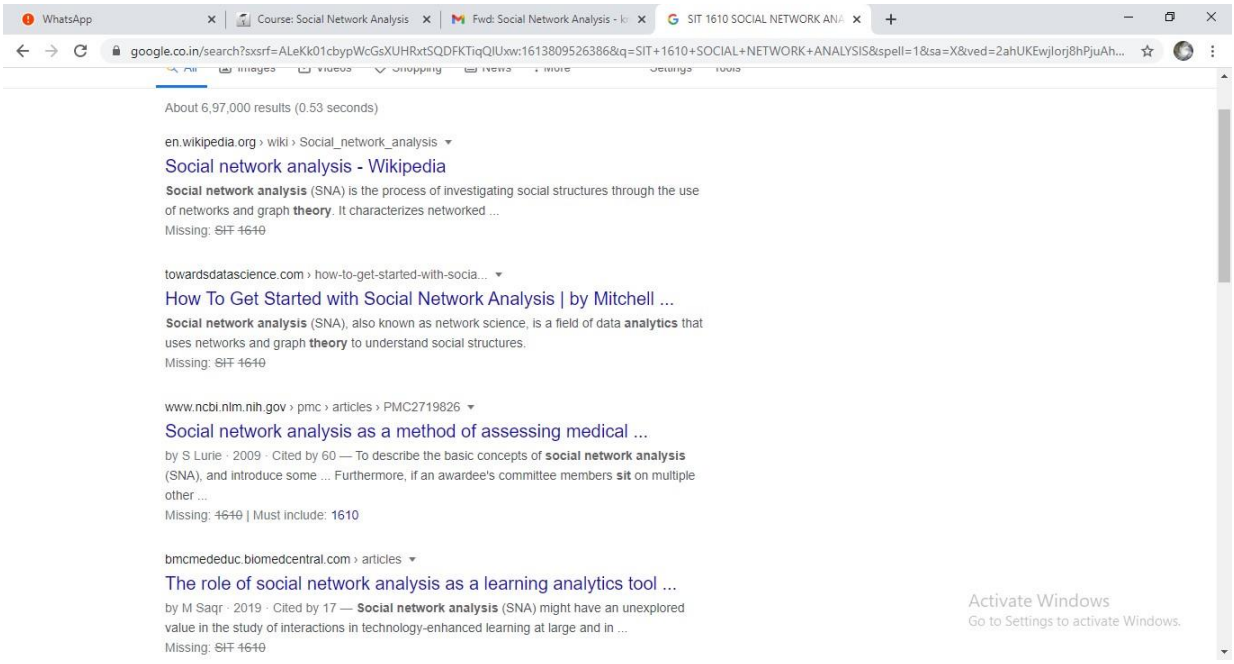
 This means adaptation to our primary interface to the vast information that constitutes the Web: the search engine.

 The following are the four questions that search engines cannot answer at the moment with satisfaction or not at all.

What's wrong with the Web?

The questions below are specific. They represent very general categories of search tasks. In each of these cases semantic technology would drastically improve the computer's ability to give more appropriate answers.

- To answer such a question using the Web one would go to the search engine and enter the most logical keyword: SIT1610 Social network analysis. The results returned by Google are shown in Figure 1
- From the top ten results only three are related to the social network analysis notes we are interested in. The word SIT1610 means a number of things. It's show the set of images, notes and general topics about Social network analysis.
- Two of the hits related to notes, three related to syllabus of social network analysis and other related to general concepts of social networking analysis.
- The problem is thus that the keyword *SIT1610 is polysemous*
- The reason is search engines know that users are not likely to look at more than the topten results. 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.
- This allows the user to quickly realize the ambiguity of the query and to make it more Specific.



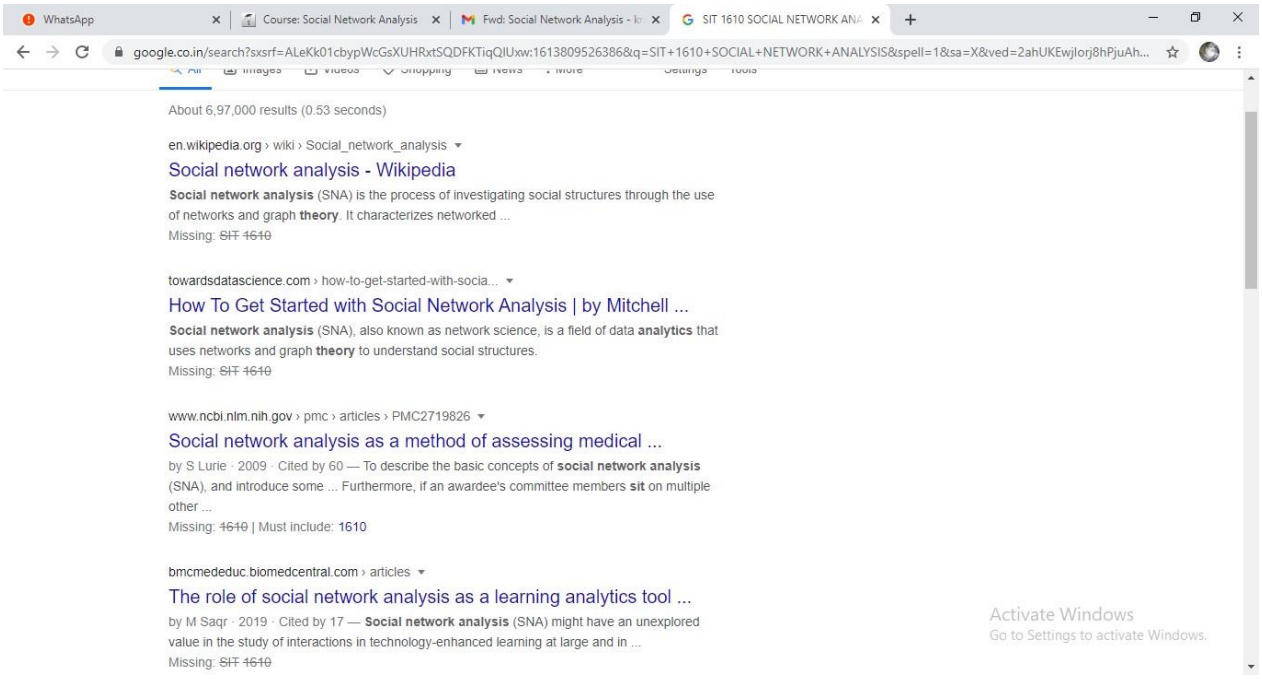


Fig.1 Search results for the keyword SIT1610 *Social network analysis* using Google

2. Show me photo of Paris

Typing "**Paris photos**" in search engine returned the result in google image as below. The search engine fails to discriminate two categories of images: i. related to the city of Paris and ii. showing Paris Hilton While the search engine does a good job with retrieving documents, the results of image searches in general are disappointing. For the keyword *Paris* most of us would expect photos of places in Paris or maps of the city. In reality only about half of the photos onthe first page, a quarter of the photos on the second page and a fifth on the third page are directly related to our concept of Paris. The rest are about clouds, people, signs, diagrams etc

Problems:

- ❖ Associating photos with keywords is a much more difficult task than simply looking for keywords in the texts of documents.
- ❖ Automatic image recognition is currently a largely unsolved research problem.
- ❖ Search engines attempt to understand the meaning of the image solely from its context

Find new music that I (might) like This is a difficult query. From the perspective of automation, music retrieval is just as problematic as image search. search engines do not exist for different reasons: most music on the internet is shared illegally through peer-to-peer systems that are completely out of reach for search engines. Music is also a fast moving good; search engines typically index the Web once a month and therefore too slow for the fast moving world of music releases. On the other hand, our musical taste might change in which case this query would need to change its form. A description of our musical taste is something that we might list on our homepage but it is not something that we would like to keep typing in again for accessing different music-related services on the internet.

Tell me about music players with a capacity of at least 4GB

This is a typical e-commerce query: looking for a product with certain characteristics.

One of the immediate concerns is that translating this query from natural language to the boolean language of search engines is (almost) impossible.

The search engine will not know that 4GB is the capacity of the music player.

Problem is that general purpose search engines do not know anything about music players or their properties and how to compare such properties.

Another bigger problem in our machines is trying to collect and aggregate product information from the Web. The information extraction methods used for this purpose have a very difficult task and it is easy to see why if we consider how a typical product description page looks like to the eyes of the computer.

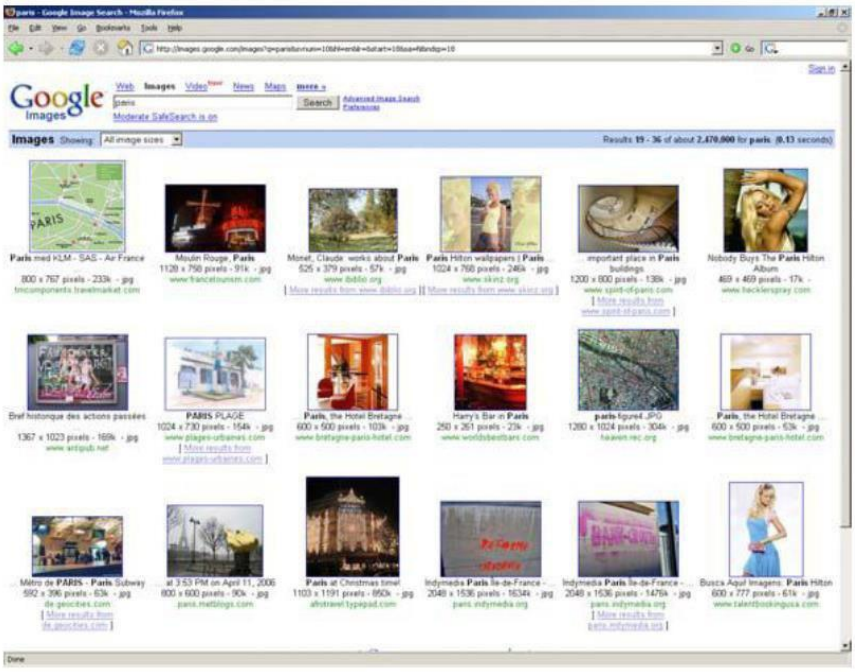
Even if an algorithm can determine that the page describes a music player, information about the product is very difficult to spot.

Further, what one vendor calls "capacity" and another may call "memory". In order to compare music players from different shops we need to determine that these two properties are actually the same and we can directly compare their values.

Google Scholar and CiteSeer are the two most well-known examples.

They suffer from the typical weaknesses of information extraction, e.g. when searching *York Sure*, the name of a Semantic Web researcher, Scholar returns also publications that are published in New York, but have otherwise nothing to do with the researcher in question. The cost of such errors is very low, however: most of us just ignore the incorrect results.

In the first case, the search is limited to the stores known by the system. On the other hand, the second method is limited by the human effort required for maintaining product categories as well as locating websites and implementing methods of information extraction. As a result, these comparison sites feature only a selected number of vendors, product types and attributes.



How to improve current Web?

- Increasing automatic linking among data
- Increasing recall and precision in search
- Increasing automation in data integration
- Increasing automation in the service life cycle

Adding semantics to data and services is the solution!

1.3 DEVELOPMENT OF SEMANTIC WEB

RESEARCH, DEVELOPMENT AND STANDARDIZATION

The vision of extending the current human-focused Web with machine process able descriptions of web content has been first formulated in 1996 by Tim Berners-Lee, the original inventor of the Web.

 The **Semantic Web** has been actively promoted by the **World Wide Web Consortium**. The organization is chiefly responsible for setting technical standards on the Web.

 The Semantic Web has quickly attracted significant interest from funding agencies on both sides of the Atlantic, reshaping much of the AI research agenda in a relatively short period of time.

 For example, Natural Language Processing and Information Retrieval have been applied to acquiring knowledge from the World Wide Web.

 As the Semantic Web is a relatively new, dynamic field of investigation, it is difficult to precisely delineate the boundaries of this network. For research on the Semantic Web

community, researchers have submitted publications or held an organizing role at any of the past International Semantic Web Conferences.

 The complete list of individuals in this community consists of 608 researchers mostly from academia (79%) and to a lesser degree from industry (21%). Geographically, the community covers much of the United States, Europe, with some activity in Japan and Australia.

 The core technology of the Semantic Web, logic-based languages for knowledge representation and reasoning has 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.

Technology adoption
The Semantic Web was originally conceptualized as an extension of the current Web, i.e. as the application of metadata for describing Web content. In this vision, the content that is already on the Web.

 This vision was soon considered to be less realistic.

 The alternative view predicted that the Semantic Web will first break through behind the scenes and not with the ordinary users, but among large providers of data and services.

 The second vision predicts that the Semantic Web will be primarily a "web of data" operated by data and service providers.

 That the Semantic Web is formulated as a vision points to the problem of bootstrapping the Semantic Web.

Difficulties:

The problem is that as a technology for developers, users of the Web never experiences the Semantic Web directly, which makes it difficult to convey Semantic Web technology to stakeholders. Further, most of the times the gains for developers are achieved over the long term, i.e. when data and services need to reused and re-purposed. The semantic web suffers from **Fax-effect**.

When the first fax machines were introduced, they came with a very hefty price tag. Yet they were almost useless. The usefulness of a fax comes from being able to communicate with other fax users. In this sense every fax unit sold increases the value of all fax machines in use.

 With the **Semantic Web** the beginning the price of technological investment is very high. One has to adapt the new technology **which requires an investment in learning**. The technology **needs time to become more reliable.**

 It required a certain kind of agreement to get the system working on a global scale: all fax machines needed to adopt the same protocol for communicating over the telephone line. This is similar to the case of the Web where global interoperability is guaranteed by the standard protocol for communication (HTTP).

 In order to exchange meaning there has to be a minimal external agreement on the meaning of some primitive symbols, i.e. on what is communicated through the network.

Our machines can also help in this task to the extent that some of the meaning can be describedin formal rules (**e.g. if A is true, B should follow**). But formal knowledge typically captures onlythe smaller part of the intended meaning and thus there needs to be a common grounding in an external reality that is shared by those at separate ends of the line.

 To follow the popularity of Semantic Web related concepts and Semantic Web standards on the Web, have **executed a set of temporal queries using the search engine Altavista**.

 The queries contained single terms plus a disambiguation term where it was necessary. Each query measured the number of documents with the given term(s) at the given point in time.

The below figure shows the number of documents with the terms *basketball*, *Computer Science*, and *XML*. The flat curve for the term basketball validates this strategy: the popularity of basketball to be roughly stable over this time period. Computer Science takes less and less share of the Web as the Web shifts from scientific use to everyday use. The share of XML, a popular pre-semantic web technology seems to grow and stabilize as it becomes a regular part of the toolkit of Web developers.

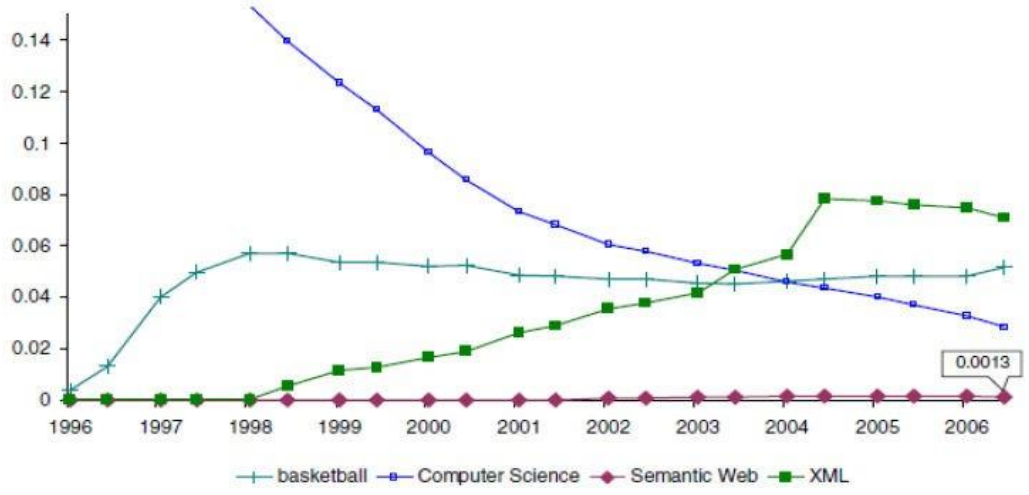


Fig2. Number of webpage with the terms *basketball*, *Computer Science*, and *XML* over time and as a fraction of the number of pages with the term *web*.

Against this general backdrop there was a look at the share of Semantic Web related terms and formats, in particular the terms *RDF*, *OWL* and the number of ontologies (Semantic Web Documents) in RDF or OWL. As **Figure 1.3.b** shows most of the curves have flattened out after January, 2004. It is not known at this point whether the dip in the share of Semantic

Web is significant. While the use of RDF has settled at a relatively high level, OWL has yet to break out from a very low trajectory.

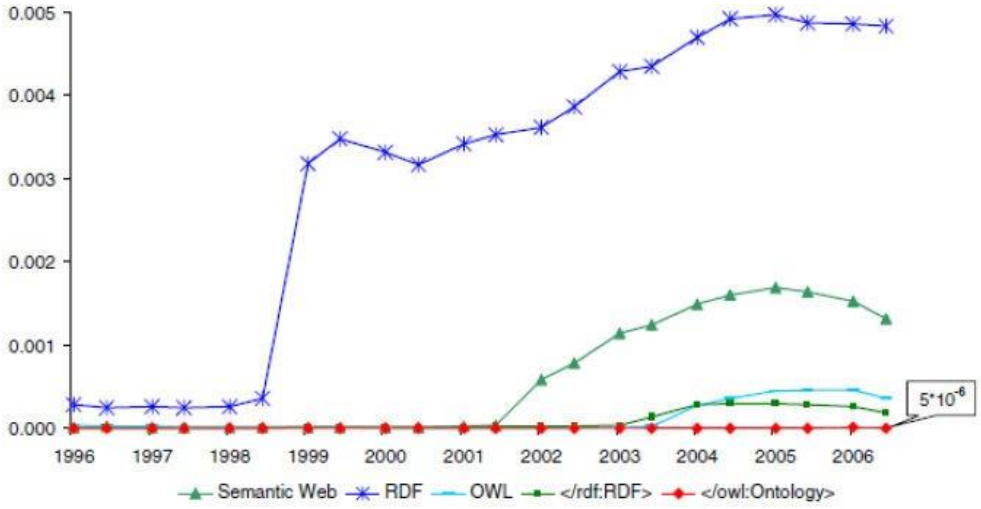


Fig3. Number of WebPages with the terms RDF, OWL and the number of ontologies in RDF or OWL over time. Again, the number is relative to the number of pages with the term web.

The share of the mentioning of Semantic Web formats versus the actual number of Semantic Web documents using that format. The resulting *talking vs. doing* curve shows the phenomenon of technology hype in both the case of XML, RDF and OWL. this is the point where the technology "makes the press" and after which its becoming increasingly used on the Web.

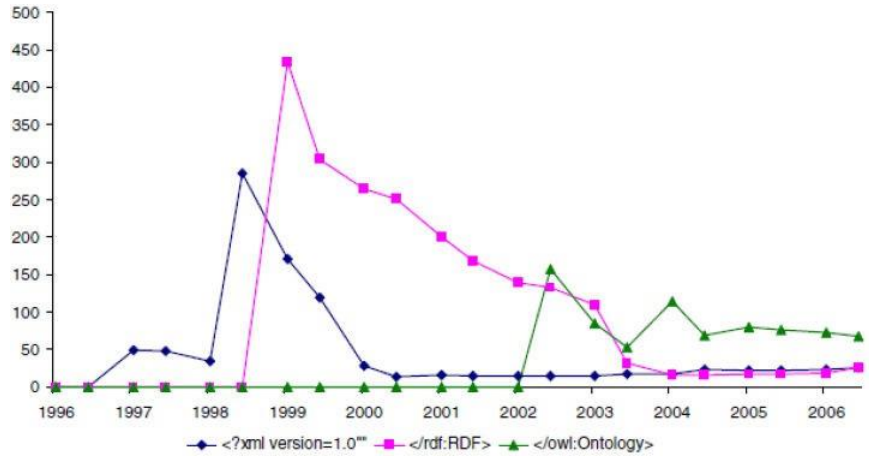


Fig.4 The hype cycle of Semantic Web related technologies as shown by the number of web pages about a given technology relative to its usage

The five-stage *hype cycle* of Gartner Research is defined as follows: The first phase of a Hype Cycle is the "technology trigger" or breakthrough, product launch or other event that generates significant press and interest. In the next phase, a frenzy of publicity typically generates over-

enthusiasm and unrealistic expectations. There may be some successful applications of a technology, but there are typically more failures. Technologies enter the "trough of disillusionment" because they fail to meet expectations and quickly become unfashionable. Although the press may have stopped covering the technology, some businesses continue through the "slope of enlightenment" and experiment to understand the benefits and practical application of the technology. A technology reaches the "plateau of productivity" as the benefits of it become widely demonstrated and accepted. The technology becomes increasingly stable andevolves in second and third generations. The final height of the plateau varies according to whether the technology is broadly applicable or benefits only a niche market.

 Although the word hype has attracted some negative connotations, hype is unavoidable for the adoption of network technologies such as the Semantic Web.

 While standardization of the Semantic Web is mostly complete, Semantic Web technology is not reaching yet the mainstream user and developer community of the Web.

In particular, the adoption of RDF is lagging behind XML, even though it provides a better alternative and thus many hoped it would replace XML over time.

 The recent support for Semantic Web standards by vendors such as Oracle23 will certainly inspire even more confidence in the corporate world. This could lead an earlier realization of the vision of the Se mantic Web as a "web of data", which could ultimately result in a resurgence of general interest on the Web.

1.4 THE EMERGENCE OF WEB

The Web was a read-only medium for a majority of users. The web of the 1990s was much like the combination of a phone book and the yellow pages and despite the connecting power of hyperlinks it instilled little sense of community among its users. This passive attitude toward the Web was broken by a series of changes in usage patterns and technology that are now referred to as Web 2.0, a buzzword coined by Tim O'Reilly.

History of web 2.0

These set of innovations in the architecture and usage patterns of the Web led to an entirely different role of the online world as a platform for intense communication and social interaction. A recent major survey based on interviews with 2200 adults shows that the internet significantly improves Americans' capacity to maintain their social networks despite early fears about the effects of diminishing real life contact.

Blogs The first wave of socialization on the Web was due to the appearance of *blogs*, *wikis* and other forms of web-based communication and collaboration. Blogs and wikis attracted mass popularity from around 2003

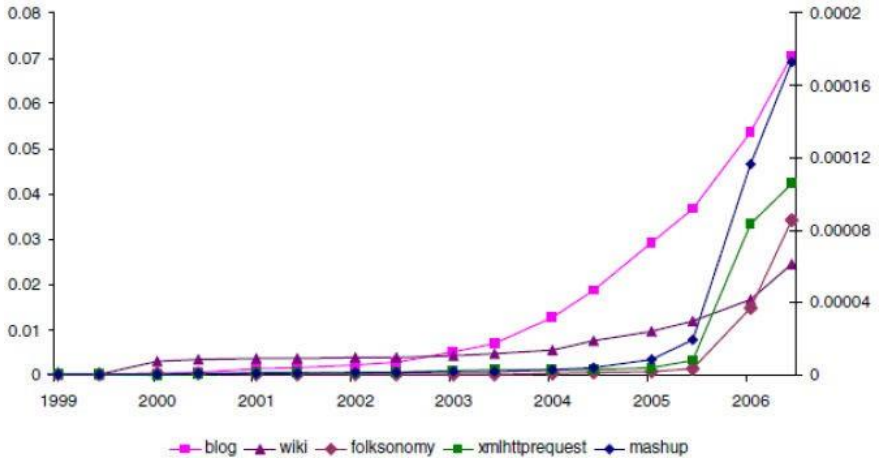


Fig.5 Development of the social web.

The fraction of web pages with the terms *blogs, wiki* over time is measured on the left vertical axis. The fraction of web pages with the terms *folk sonomy, XmlHttpRequest* and *mashup* is measured on the right hand vertical axis.

For adding content to the Web: editing blogs and wikis did not require any knowledge of HTML any more. Blogs and wikis allowed individuals and groups to claim their personal space on the Web and fill it with content at relative ease. Even more importantly, despite that weblogs have been first assessed as purely personal publishing (similar to diaries), nowadays the blogosphere is widely recognized as a densely interconnected social network through which news, ideas and influences travel rapidly as bloggers reference and reflect on each other's postings.

Example: Wikipedia, the online encyclopedia The significance of instant messaging (ICQ) is also not just instant communication (phone is instantaneous, and email is almost instantaneous), but the ability to see who is online, a transparency that induces a sense of social responsibility.

Social networks

The first *online social networks* also referred to as social networking services. It entered the field at the same time as blogging and wikis started to take off. Attracted over five million registered users followed by Google and Microsoft. These sites allow users to post a profile with basic information, to invite others to register and to link to the profiles of their friends. The system also makes it possible to visualize and browse the resulting network in order to discover friends in common, friends thought to be lost or potential new friendships based on shared interests.

The latest services are thus using user profiles and networks to stimulate different exchanges: photos are shared in Flickr, bookmarks are exchanged in del.icio.us, plans and goals unite members at 43Things. The idea of **network based exchange** is based on the **sociological observation** that social interaction creates similarity and vice versa, interaction creates similarity: friends are likely to have acquired or develop similar interests.

User profiles

Explicit user profiles make it possible for these systems to introduce rating mechanism whereby either the users or their contributions are ranked according to usefulness or trustworthiness. Ratings are explicit forms of social capital that regulate exchanges in online communities such that reputation moderates exchanges in the real world. In terms of implementation, the new web sites are relying on new ways of applying some of the pre-existent technologies. Asynchronous JavaScript and XML, or *AJAX*, which drives many of the latest websites is merely a mix of technologies that have been supported by browsers for years. User friendliness is a preference for formats, languages and protocols that are easy to use and develop with, in particular script languages, formats such as JSON, protocols such as REST.

This is to support rapid development and prototyping. For example: flickr Also, borrowing much of the ideology of the open source software movement, Web 2.0 applications open up their data and services for user experimentation: Google, Yahoo and countless smaller web sites. through lightweight APIs content providers do the same with information in the form of RSS feeds. The results of user experimentation with combinations of technologies are the so-called *mashups*. Mashups is a websites based on combinations of data and services provided by others. The best example of this development are the mashups based on Google's mapping service such as HousingMaps.

Web 2.0 + Semantic Web =Web 3.0?

Web 2.0 is often contrasted to the Semantic Web. the ideas of Web 2.0 and the Semantic Web are not exclusive alternatives: while Web 2.0 mostly effects how users interact with the Web, while the Semantic Web opens new technological opportunities for web developers in combiningdata and services from different sources.

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.***

It addresses a primary concern of the Semantic Web community, namely whether users would be willing to provide metadata to bootstrap the Semantic Web. The Semantic Web was originally also expected to be filled by users annotating Web resources, describing their home pages and multimedia content.

It seems clear that many are in fact willing to provide structured information, provided that they can do so in a task oriented way and through a user-friendly interface that hides the complexity of the underlying representation. Micro formats, for example, proved to be more popular due to the easier authoring using existing HTML attributes.

Web pages created automatically from a database (such as blog pages or personal profile pages) can encode metadata in micro formats without the user necessarily being aware of it. For example, blog search engines are able to provide search on the properties of the author or the news item.

Noting this, the idea of providing ways to encode RDF into HTML pages has resurfaced. There are also works under way to extend the MediaWiki software behind Wikipedia to allow

users to encode facts in the text of articles while writing the text. This additional, machine processable markup of facts would enable to easily extract, query and aggregate the knowledge of Wikipedia.

Similar works on entirely new Wiki systems that combine free-text authoring with the collaborative editing of structured information.

Information about the choices, preferences, tastes and social networks of users means that the new breed of applications are able to build on a much richer user profiles. Clearly, semantic technology can help in matching users with similar interests as well as matching users with available content.

- Lastly, in terms of technology what the Semantic Web can offer to the Web 2.0 community is a standard infrastructure for the building creative combinations of data and services. Standard formats for exchanging data and schema information, support for data integration, along with standard query languages and protocols for querying remote data sources provide a platform for the easy development of mashups.

1.5 STATISTICAL PROPERTIES OF SOCIAL NETWORKS

1.6 NETWROK ANALYSIS

Social Network Analysis (SNA) is the study of social relations among a set of actors. The key difference between network analysis and other approaches to social science is the focus on relationships between actors rather than the attributes of individual actors. Network analysis takes a global view on social structures based on the belief that types and patterns of relationships emerge from individual connectivity and that the presence (or absence) of such types and patterns have substantial effects on the network and its constituents. In particular, the network structure provides opportunities and imposes constraints on the individual actors by determining the transfer or flow of resources (material or immaterial) across the network.

The focus on relationships as opposed to actors can be easily understood by an example. When trying to predict the performance of individuals in a scientific community by some measure (say, number of publications), a traditional social science approach would dictate to look at the attributes of the researchers such as the amount of grants they attract, their age, the size of the team they belong to etc. A statistical analysis would then proceed by trying to relate these attributes to the outcome variable, i.e. the number of publications. In the same context, a networkanalysis study would focus on the interdependencies within the research community.

For example, one would look at the patterns of relationships that scientists have and the potential benefits or constraints such relationships may impose on their work. For example, one may hypothesize that certain kinds of relationships arranged in a certain pattern may be beneficial to performance compared to the case when that pattern is not present. The patterns of relationships may not only be used to explain individual performance but also to hypothesize their impact on the network itself (network evolution). Attributes typically play a secondary role in network studies as control variables.1 SNA is thus a different approach to social phenomena and therefore requires a new set of concepts and new methods for data collection and analysis.

Network analysis provides a vocabulary for describing social structures, provides formal models that capture the common properties of all (social) networks and a set of methods applicable to the analysis of networks in general. The concepts and methods of network analysis are grounded in a formal description of networks as graphs.

Methods of analysis primarily originate from graph theory as these are applied to the graph representation of social network data. (Network analysis also applies statistical and probabilistic methods and to a lesser extent algebraic techniques.) It is interesting to note that the formalization of network analysis has brought much of the same advantages that the formalization of knowledge on the Web (the Semantic Web) is expected to bring to many application domains. Previously vaguely defined concepts such as social role or social group could now be defined on a formal model of networks, allowing to carry out more precise discussions in the literature and to compare results across studies.

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. Often records of social interaction (publication databases, meeting notes, newspaper articles, documents and databases of different sorts) are used to build a model of social networks

1.7 DEVELOPMENT OF SOCIAL NETWORK ANALYSIS

The field of Social Network Analysis today is the result of the convergence of several streams of applied research in sociology, social psychology and anthropology. Many of the concepts of network analysis have been developed independently by various researchers often through empirical studies of various social settings.

For example, many social psychologists of the 1940s found a formal description of social groups useful in depicting communication channels in the group when trying to explain processes of group communication. Already in the mid-1950s anthropologists have found network representations useful in generalizing actual field observations, for example when comparing the level of reciprocity in marriage and other social exchanges across different cultures.

Some of the concepts of network analysis have come naturally from social studies. In an influential early study at the Hawthorne works in Chicago, researchers from Harvard looked at the workgroup behavior (e.g. communication, friendships, helping, controversy) at a specific partof the factory, the bank wiring room. The investigators noticed that workers themselves used specific terms to describe who is in “our group”.

The researchers tried to understand how such terms arise by reproducing in a visual way the group structure of the organization as it emerged from the individual relationships of the factory workers.

2. In another study of mixed-race city in the Southern US researchers looked at the network of overlapping “cliques” defined by race and age.

3. They also went further than the Hawthorne study in generating hypotheses about the possible connections between cliques.

Despite the various efforts, each of the early studies used a different set of concepts and different methods of representation and analysis of social networks. However, from the 1950s network analysis began to converge around the unique world view that distinguishes network analysis from other approaches to sociological research.

This convergence was facilitated by the adoption of a graph representation of social networks usually credited to Moreno. What Moreno called a sociogram was a visual representation of social networks as a set of nodes connected by directed links. The nodes represented individuals in Moreno's work, while the edges stood for personal relations. However, similar representationscan be used to depict a set of relationships between any kind of social unit such as groups, organizations, nations etc. While 2D and 3D visual modeling is still an important technique of network analysis, the sociogram is honored mostly for opening the way to a formal treatment of network analysis based on graph theory.

The following decades have seen a tremendous increase in the capabilities of network analysis mostly through new applications. SNA gains its relevance from applications and these settings in turn provide the theories to be tested and greatly influence the development of the methods and the interpretation of the outcomes. For example, one of the relatively new areas of network analysis is the analysis of networks in entrepreneurship, an active area of research that builds andcontributes to organization and management science.

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 is the advances in dealing with longitudinal data. New probabilistic models are capable of modeling 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 increasing variety of applications and related advances in methodology can be best observed at the yearly Sunbelt Social Networks Conference series, which started in 1980.

4. The field of Social Network Analysis also has a journal of the same name since 1978, dedicated largely to methodological issues.

5. However, articles describing various applications of social network analysis can be found in almost any field where networks and relational data play an important role.

While the field of network analysis has been growing steadily from the beginning, there have been two developments in the last two decades that led to an explosion in network literature. First, advances in information technology brought a wealth of electronic data and significantly increased analytical power.

Second, the methods of SNA are increasingly applied to networks other than social networks such as the hyperlink structure on the Web or the electric grid. This advancement —brought forward primarily by physicists and other natural scientists— is based on the discovery that many networks in nature share a number of commonalities with social networks.

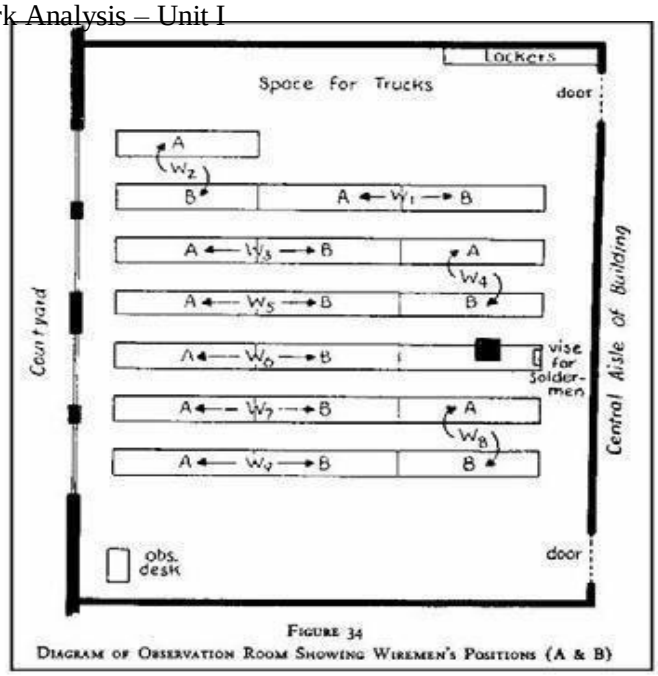


FIGURE 34
DIAGRAM OF OBSERVATION ROOM SHOWING WIREMEN'S POSITIONS (A & B)

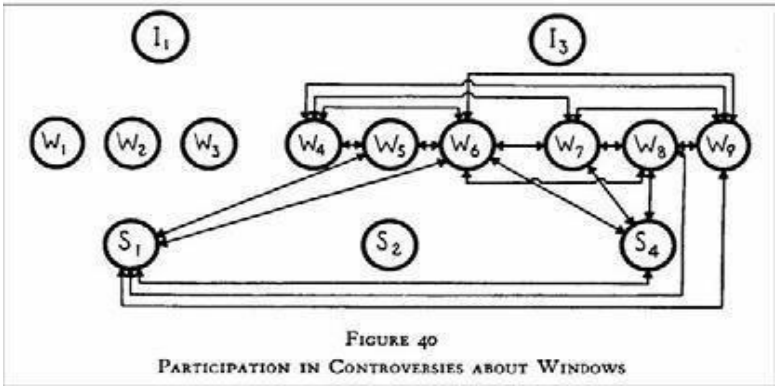


FIGURE 40
PARTICIPATION IN CONTROVERSIES ABOUT WINDOWS

In the following, we will also talk about networks in general, but it should be clear from the text that many of the measures in network analysis can only be strictly interpreted in the context of social networks or have very different interpretation in networks of other kinds.

Fig.6 The upper part shows the location of the workers in the wiring room, while the lower part is a network image of fights about the windows between workers (W), solderers (S) and inspectors (I).

The term **socialnetwork**"" has been introduced by Barnes in 1954. This convergence was facilitated by the adoption of a graph representation of social networks called as

Sociogram usually credited to Moreno.

Sociogram was a visual representation of social networks as a set of nodes connected by directed **links**. The **nodes** represented individuals while the edges stood for **personal relations**. The sociogram is honored mostly for opening the way to a formal treatment of network analysis based on graph theory.

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 modeling 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.

While the field of network analysis has been growing steadily from the beginning, there have been two developments in the last two decades that led to an explosion in network literature

First, advances in information technology brought a wealth of electronic data and significantly increased analytical power.

Second, the methods of SNA are increasingly applied to networks other than social networks such as the hyperlink structure on the Web or the electric grid

This advancement is based on the discovery that many networks in nature share a number of commonalities with social networks.

1.8 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.

1.8.1 The global structure of networks

A Social network can be represented as a Graph G = (V,E) where V denotes finite set of vertices and E denoted finite set of Edges.
Each graph can be associated with its characteristic matrix $M: =(m_{i,j})_{n*n}$ where n =|V|

$$m_{i,j} = \left\{ \begin{array}{l|l} 1 & (v_i, v_j) \in E \\ 0 & otherwise \end{array} \right\},$$

A component is a maximal connected subgraph. Two vertices are in the same (strong) component if and only if there exists a (directed) path between them.

American psychologist Stanley Milgram experiment about the structure of social networks. Milgram calculated the average of the length of the chains and concluded that the experiment showed that on average Americans are no more than six steps apart from each other. While this is also the source of the expression *six degrees of separation* the actual number is rather dubious:

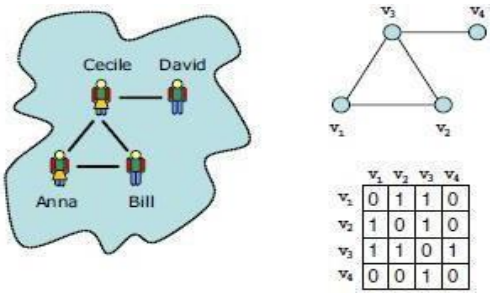


Fig. 7 Most network analysis methods work on an abstract, graph based representation of real world networks.

Formally, what Milgram estimated is the size of the average shortest path of the network, which is also called *characteristic path length*. The shortest path between two vertices v_s and v_t is a paththat begins at the vertex v_s and ends in the vertex v_t and contains the least possible number of vertices. The shortest path between two vertices is also called a *geodesic*. The longest geodesic in the graph is called the diameter of the graph: this is the maximum number of steps that is required between any two nodes. The average shortest path is the average of the length of the geodesics between all pairs of vertices in the graph.

A practical impact of Milgram's finding structures is as that possible models for social networks. The two dimensional lattice model shown in Figure.

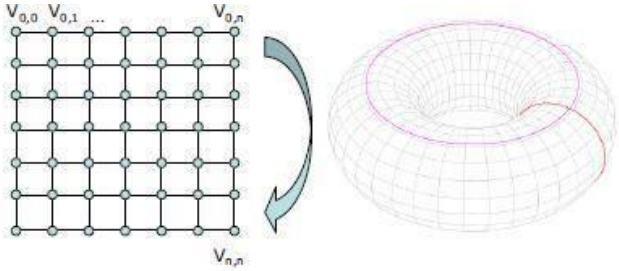
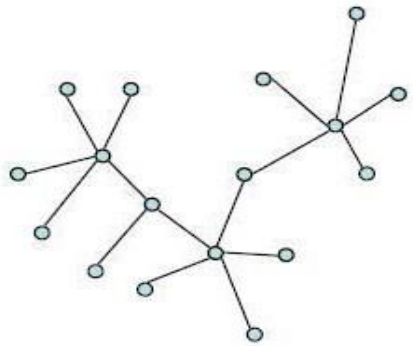


Fig.8 The 2D lattice model of networks (left). By connecting the nodes on the opposite borders of the lattice we get a toroidal lattice (right).

Clustering for a single vertex can be measured by the actual number of the edges between the neighbors of a vertex divided by the possible number of edges between the neighbors. When taken the average over all vertices we get to the measure known as *clustering coefficient*. The clustering coefficient of 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.

Fig.9 A tree is a connected graph where there are no loops and paths leading from a vertex to itself.

The macro-structure of social networks

The image that emerges is one of dense clusters or social groups sparsely connected to each other by a few ties as shown in Figure 1.7.d. For example, this is the image that appears if we investigate the co-authorship networks of a scientific community. Bounded by limitations of space and resources, scientists mostly co-operate with colleagues from the same institute. Occasional exchanges and projects with researchers from abroad, however, create the kind of shortcut ties that Watts explicitly incorporated within his model. These shortcuts make it possible for scientists to reach each other in a relatively short number of steps.

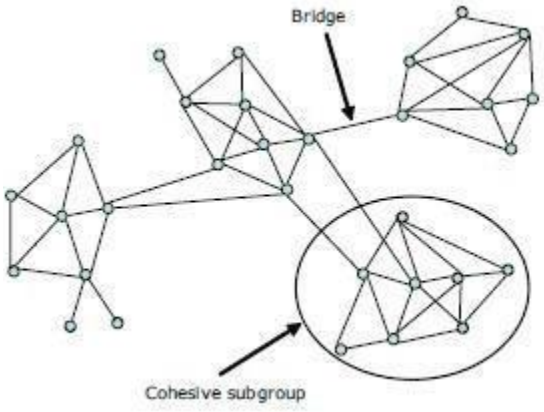


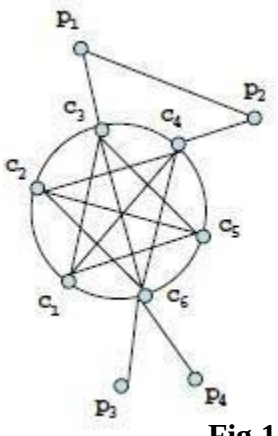
Fig.10 Most real world networks show a structure where densely connected subgroups are linked together by relatively few bridges

Clustering a graph into subgroups allows us to visualize the connectivity at a group level. ***Core-Periphery (C/P) 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 (see Figure 1.7. e). The matrix form of a core periphery structure is a

$\begin{pmatrix} 1 & . \\ . & 0 \end{pmatrix}$ matrix

The result of the optimization is a classification of the nodes as core or periphery and a measure of the error of the solution.



c_1	c_2	c_3	c_4	c_5	c_6	p_1	p_2	p_3	p_4
1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	0	0	0
1	1	1	1	1	1	0	1	0	0
1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	0	0	1	1
0	0	1	0	0	0	0	1	0	0
0	0	0	1	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0

Fig.11

The **structural dimension** of social capital refers to patterns of relationships or positions that provide benefits in terms of accessing large, important parts of the network.

Degree centrality equals the graph theoretic measure of degree, i.e. the number of (incoming, outgoing or all) links of a node.

Closeness centrality, which is obtained by calculating the average (geodesic) distance of a node to all other nodes in the network. In larger networks it makes sense to constrain the size of the neighborhood in which to measure closeness centrality. It makes little sense, for example, to talk about the most central node on the level of a society. The resulting measure is called ***local closeness centrality*.**

Two other measures of power and influence through networks are ***broker positions*** and ***weak ties.***

Betweenness is defined as the proportion of paths — among the geodesics between all pairs of nodes—that pass through a given actor.

A **structural hole** occurs in the space that exists between closely clustered communities.

Lastly, he proves that the structural holes measure correlates with creativity by establishing a linear equation between the network measure and the individual characteristics on one side of the equation and creativity on the other side.

1.9 DISCUSSION NETWORKS

One of the foremost studies to illustrate the versatility of electronic data is a series of works from the Information Dynamics Labs of Hewlett-Packard. Tyler, Wilkinson and Huber man analyze communication among employees of their own lab by using the corporate email archive. They recreate the actual discussion networks in the organization by drawing a tie between two individuals if they had exchanged at least a minimum number of total emails in a given period, filtering out one-way relationships.

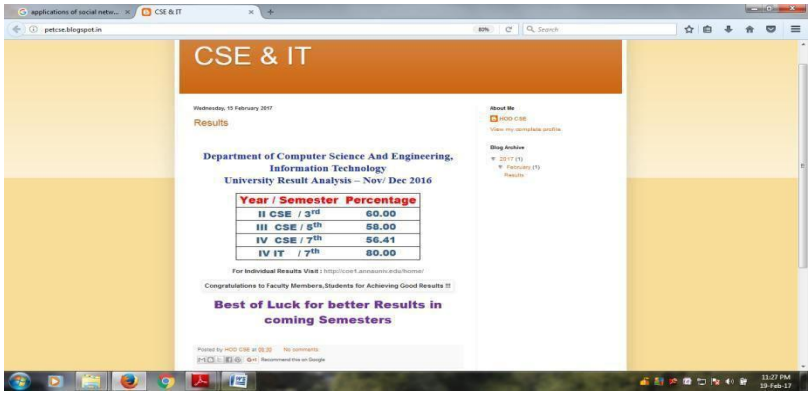
The studies of electronic communication networks based on email data are limited by privacy concerns. For example, in the HP case the content of messages had to be ignored by the researchers and the data set could not be shared with the community.

Public forums and mailing lists can be analyzed without similar concerns. The W3C — which is also the organization responsible for the standardization of Semantic Web technologies—is unique among standardization bodies in its commitment to transparency toward the general public of the Internet and part of this commitment is the openness of the discussions within the working groups.

1.10 BLOGS AND ONLINE COMMUNITIES

Content analysis has also been the most commonly used tool in the computer aided analysis of blogs (web logs), primarily with the intention of trend analysis for the purposes of marketing. While blogs are often considered as "person themselves know that blogs are much more than that: modern blogging tools allow to easily comment and react to the comments of other bloggers, resulting in webs of communication among bloggers.

These discussion networks also lead to the establishment of dynamic communities, which often manifest themselves through syndicated blogs (aggregated blogs that collect posts from a set of authors blogging on similar topics), blog rolls (lists of discussion partners on a personal blog) and even result in real world meetings such as the Blog Walk series of meetings.



Link to Other Blog

Link to Another Blog Post

Links from Other Blogs

Comments

Word Press. Yes, there are other blogging platforms and some of them may be easier for new

computer users and non-techies to use. ...

Gmail. I have many email addresses which all automatically send to my Gmail email account. ...

Google Analytics. ...

MailChimp. ...

Evernote. ...

My Hours. ...

Rapportive. ...

Dropbox

The 2004 US election campaign represented a turning point in blog research as it has been the first major electoral contest where blogs have been exploited as a method of building networks among individual activists and supporters. Blog analysis has suddenly shed its image as relevant only to marketers interested in understanding product choices of young demographics; following this campaign there has been explosion in research on the capacity of web logs for creating and maintaining stable, long distance social networks of different kinds.

Online community spaces and social networking services such as MySpace, LiveJournal cater to socialization even more directly than blogs with features such as social networking (maintaining lists of friends, joining groups), messaging and photo sharing.4 As they are typically used by a much younger demographic they offer an excellent opportunity for studying changes in youth culture.

1.11 WEB BASED NETWORKS

There are two features of web pages that are considered as the basis of extracting social relations: **links and co-occurrences**.

The **linking structure** of the Web is considered as proxy for real world relationships as links are chosen by the author of the page and connect to other information sources that are considered authoritative and relevant enough to be mentioned.

The biggest drawback of this approach is that such direct links between personal pages are very sparse: due to the increasing size of the Web searching has taken over browsing as the primary mode of navigation on the Web.

As a result, most individuals put little effort in creating new links and updating link targets or have given up linking to other personal pages altogether.

Co-occurrences of names in web pages can also be taken as evidence of relationships and are a

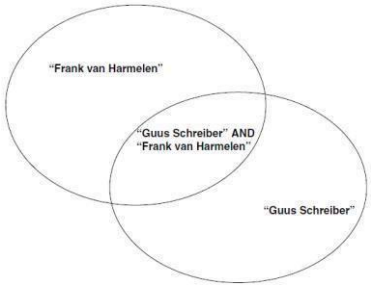
more frequent phenomenon.

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.

Web mining is the application of text mining to the content of web pages. The techniques employed here are statistical methods possibly combined with an analysis of the contents of web pages.

Using the search engine Altavista the system collected page counts for the individual names as well as the number of pages where the names co-occurred.

Note that this corresponds to a very shallow parsing of the web page as indirect references are not counted this way (e.g. the term "the pre with George Bush even if he was mentioned as the president elsewhere in the text.)



Tie strength was calculated by dividing the number of co-occurrences with the number of pages returned for the two names individually (see Figure).

Also known as the Jaccard-coefficient, this is basically the ratio of the sizes of two sets: the intersection of the sets of pages and their union.

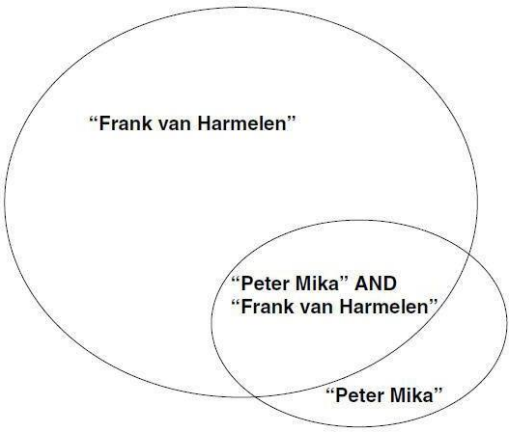
The resulting value of tie strength is a number between zero (no co-occurrences) and one (no separate mentioning, only co-occurrences). If this number has exceeded a certain fixed threshold it was taken as evidence for the existence of a tie.

The number of pages that can be found for the given individuals or combination of individuals.

The reason is that the Jaccard-coefficient is a relative measure of co-occurrence and it does not take into account the absolute sizes of the sets. In case the absolute sizes are very low we can easily get spurious results.

A disadvantage of the Jaccard-coefficient is that it penalizes ties between an individual whose name often occurs on theWeb and less popular individuals (see Figure 3.4).

In the science domain this makes it hard to detect, for example, the ties between famous professors and their PhD students. In this case while the name of the professor is likely to occur on a large percentage of the pages of where the name of the PhD student occurs but not vice versa.



For this reason we use an asymmetric variant of the coefficient. In particular, we divide the number of pages for the individual with the number of pages for both names and take it as evidence of a directed tie if this number reaches a certain threshold.

Semantic Similarity-Based Clustering of Web Documents Using Fuzzy C-Means. International Journal of Computational Intelligence and Applications 14(3) (2015) 2013

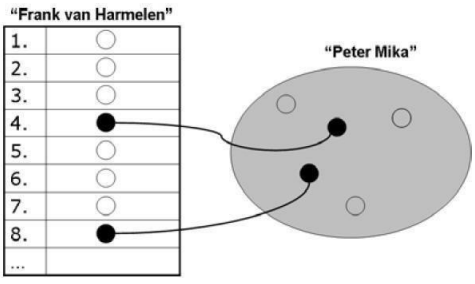
A Hybrid Approach Using PSO and K-Means for Semantic Clustering of Web Documents. J. Web Eng. 12(3&4): 249-264 (2013)

Associate researchers with topics in a slightly different way. The system calculates the strength of association between the name of a given person and a certain topic.

There have been several approaches to deal with name ambiguity. Instead of a single name they assume to have a list of names related to each other. They disambiguate the appearances by clustering the combined results returned by the search engine for the individual names. The clustering can be based on various networks between the returned webpages, e.g. based on hyperlinks between the pages, common links or similarity in content.

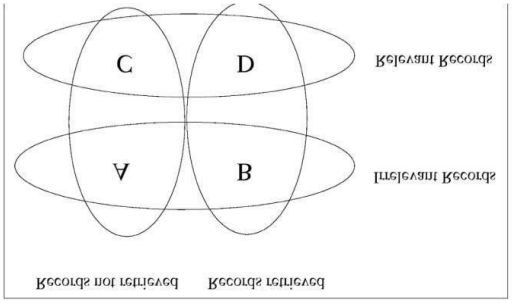
The idea is that such key phrases can be added to the search query to reduce the set of results to those related to the given target individual.

When computing the weight of a directed link between two persons.



We consider an ordered list of pages for the first person and a set of pages for the second (the relevant set) as shown in Figure:

There are four different sets: The records which were retrieved, the records which were not retrieved, the relevant records and the irrelevant records (as annotated in the test set). The intersections of these sets (A,B,C,D) represent the following: A is the number of



irrelevant records not retrieved (true negatives), B is the number of irrelevant records retrieved(false positives), C is the number of relevant records not retrieved (false negatives) and D is the number of relevant records retrieved (true positives). Recall is defined as: Recall = TP / (TP + F N)

Precision is defined as: Precision = TP / (TP + FP)

We ask the search engine for the top N pages for both persons but in the case of the second person the order is irrelevant() as the relevance for at the position compue t n, where rel(n) is 1 if the document at position n is the relevant set and zero otherwise ($1 \leq n \leq$ N).

$$P(n) = \frac{\sum_{r=1}^{n} rel(r)}{n} \qquad P_{ave} = \frac{\sum_{r=1}^{N} P(r) * rel(r)}{N}$$

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.

This strength is determined by taking the number of the pages where the name of an interest and the name of a person co-occur divided by the total number of pages about the person.

Assign the expertise to an individual if this value is at least one standard deviation higher than the mean of the values obtained for the same concept.

The biggest technical challenge in social network mining is the disambiguation of person names

Persons names exhibit the same problems of polysemy and synonymy that we have seen in the

general case of web search. Queries for researchers who commonly use different variations of their name (e.g. Jim Hendler vs. James Hendler).

Polysemy is the association of one word with two or more distinct meanings. A polyseme is a word or phrase with multiple meanings. In contrast, a one-to-one match between a word and a meaning is called monosemy. According to some estimates, more than 40% of English words have more than one meaning.The semantic qualities or sense relations that exist between words with closely related meanings is Synonymy.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT – II – SOCIAL NETWORK ANALYSIS – SITA3005

UNIT II

UNIT 2 SOCIAL NETWORK ANALYSIS SOFTWARE, TOOLS AND LIBRARIES

Modelling and aggregating social network data: Ontological representation of social individuals — Ontological representation of social relationships - Aggregating and reasoning with social network data – Advanced representations. Social network analysis software - Tools - Libraries .

2.1 MODELLING AND AGGREGATING SOCIAL NETWORK DATA

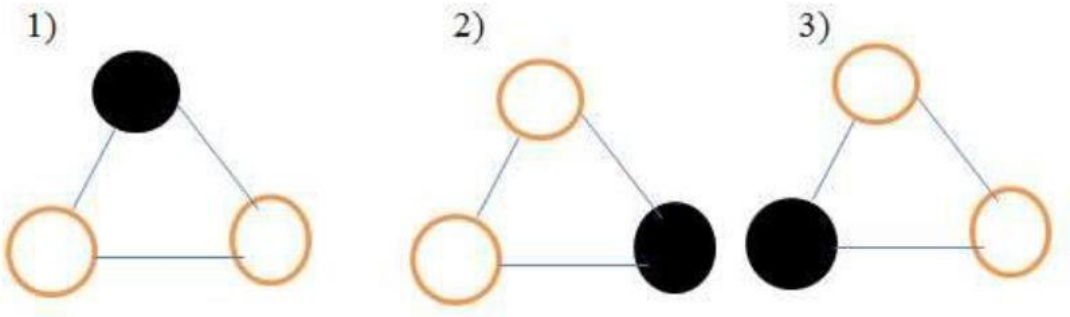
- The most common kind of social network data can be modeled by a graph where the nodes represent individuals and the edges represent binary social relationships. (Less commonly, higher-arity relationships may be represented using hyper-edges, i.e. edges connecting multiple nodes.)
- Additionally, social network studies build on attributes of nodes and edges, which can be formalized as functions operating on nodes or edges.
- A number of different, proprietary formats exist for serializing such graphs and attribute data in machine-processable electronic documents.
- The most commonly encountered formats are those used by the popular network analysis packages **Pajek and UCINET**. These are text-based formats which have been designed in

  a way so that they can be easily edited using simple text editors.
- Unfortunately, the two formats are incompatible. Further, researchers in the social sciences often represent their data initially using Microsoft Excel spreadsheets, which can be exported in the simple CSV (Comma Separated Values) format.
- The 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 advantage that GraphML files can be edited, stored, queried, transformed etc. using generic XML tools.
- Common to all these generic graph representations is that they focus on the graph structure, which is the primary input to network analysis and visualization.
- Attribute data when entered electronic form is typically stored separately from network data in Excel sheets, databases or SPSS tables.

2.2 RANDOM WALKS AND THEIR APPLICATIONS

A Random Walk in synthesis:

- Given an indirected graph and a starting point, select a neighbour at random
- Move to the selected neighbour and repeat the same process till a termination condition isverified
- The random sequence of points selected in this way is a random walk of the graph



Important parameters of random walk:

- **Access time or hitting time:** Hij is the expected number of steps before node j is visited,starting from node i
- **Commute time**: i j i: Hij + Hji
- **Cover time**: Starting from a node/distribution the expected number of steps to reachevery node.

Applications of Random Walks on Graphs

- Ranking Web Pages
- HITS on citation network
- Clustering using random walk

2.3 USE OF HADOOP AND MAP REDUCE

Map reduce

Data-parallel programming model for clusters of commodity machines

Pioneered by Google

- Processes 20 PB of data per day

Popularized by open-source Hadoop project

- Used by Yahoo!, Facebook, Amazon, …

Map Reduce used for

At Google:

1. Index building for Google Search
2. Article clustering for Google News
3. Statistical machine translation

At Yahoo!:

1. Index building for Yahoo! Search
2. Spam detection for Yahoo! Mail

At Facebook:

1. Data mining
2. Ad optimization
3. Spam detection

In research:

Analyzing Wikipedia conflicts (PARC)
Natural language processing (CMU)
Bioinformatics (Maryland)
Particle physics (Nebraska)
Ocean climate simulation (Washington)

Map Reduce Goals

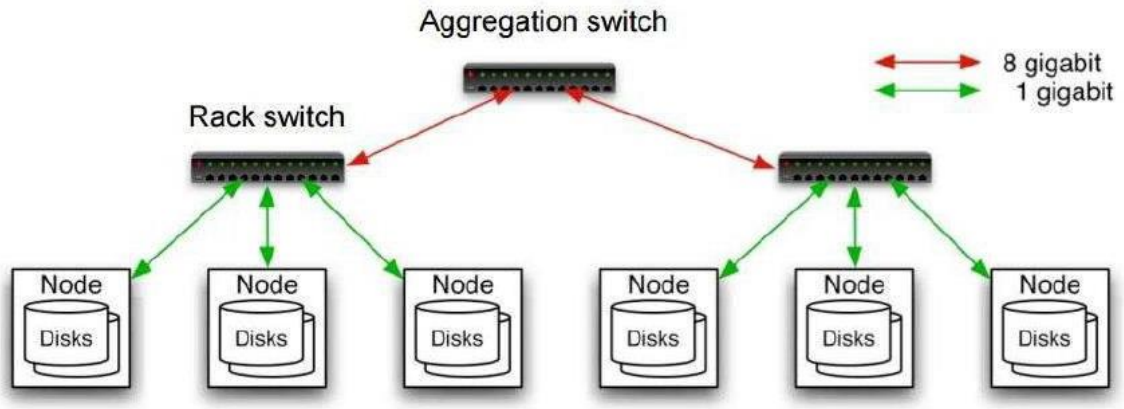
1. Scalability to large data volumes:

Scan 100 TB on 1 node @ 50 MB/s = 24 days
Scan on 1000-node cluster = 35 minutes

2. Cost-efficiency:

Commodity nodes (cheap, but unreliable)
Commodity network
Automatic fault-tolerance (fewer admins)
Easy to use (fewer programmers)

TYPICAL HADOOP CLUSTER:



40 nodes/rack, 1000-4000 nodes in cluster
1 GBps bandwidth in rack, 8 GBps out of rack

Node specs (Yahoo! terasort): 8 x 2.0 GHz cores, 8 GB RAM, 4 disks (= 4 TB?)

Challenges

Cheap nodes fail, especially if you have many
- Mean time between failures for 1 node = 3 years
- MTBF for 1000 nodes = 1 day
- Solution: Build fault-tolerance into system

Commodity network = low bandwidth
- Solution: Push computation to the data

Programming distributed systems is hard
- Solution: Users write data-parallel "map" and "reduce" functions, system handleswork distribution and faults

Hadoop Components:

Distributed file system (HDFS)
- Single namespace for entire cluster
- Replicates data 3x for fault-tolerance

MapReduce framework
- Executes user jobs specified as "map" and "reduce" functions
- Manages work distribution & fault-tolerance

Hadoop Distributed File System:

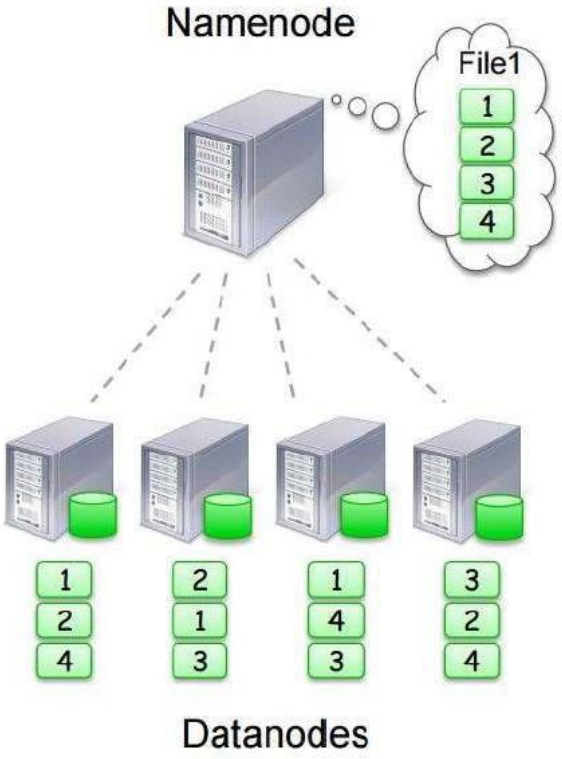
Files split into 128MB blocks

Blocks replicated across several data nodes (usually 3)

Namenode stores metadata (file names, locations, etc)

Optimized for large files, sequential reads

Files are append-only



MapReduce Programming Model:

- Data type: key-value records
- Map function:

$$(K_{in}, V_{in}) \rightarrow list(K_{inter}, V_{inter})$$

- Reduce function:

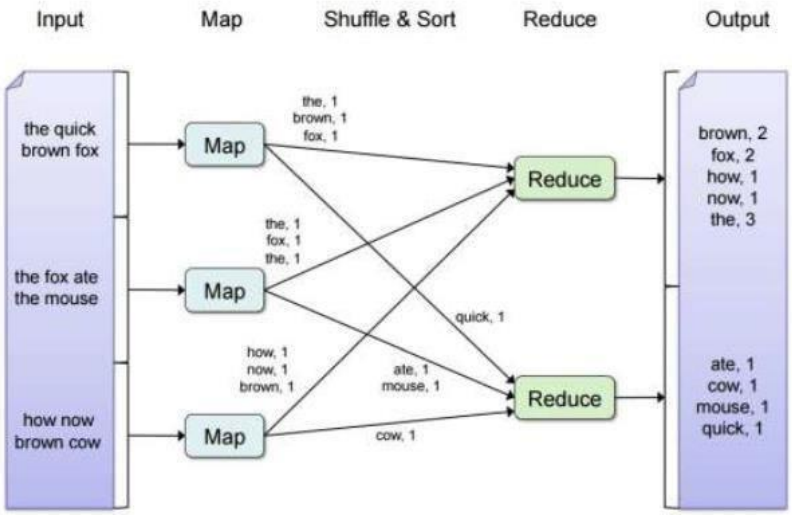
$$(K_{inter}, list(V_{inter})) \rightarrow list(K_{out}, V_{out})$$

Example: Word Count:

```
def mapper(line):
  foreach word in line.split():
    output(word, 1)

def reducer(key, values):
  output(key, sum(values))
```

Word Count Execution:



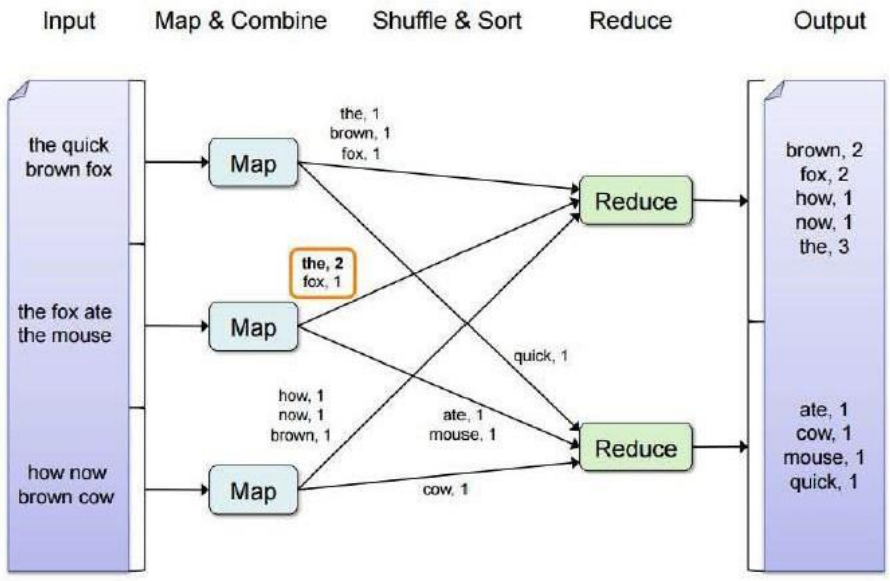
An Optimization: The Combiner

- Local aggregation function for repeated keys produced by same map

• For associative ops. like sum, count, max

• Decreases size of intermediate data

Word Count with Combiner



MapReduce Execution Details:

• Mappers preferentially placed on same node or same rack as their input block

  – Push computation to data, minimize network use

• Mappers save outputs to local disk before serving to reducers

  – Allows having more reducers than nodes

  – Allows recovery if a reducer crashes

Fault Tolerance in MapReduce:

If a task crashes:

– Retry on another node

• OK for a map because it had no dependencies

• OK for reduce because map outputs are on disk

– If the same task repeatedly fails, fail the job or ignore that input block

If a node crashes:

– Relaunch its current tasks on other nodes

– Relaunch any maps the node previously ran

• Necessary because their output files were lost along with the crashed node

If a task is going slowly (straggler):

– Launch second copy of task on another node

– Take the output of whichever copy finishes first, and kill the other one

2.4 ONTOLOGICAL REPRESENTATION OF SOCIAL INDIVIDUALSAND RELATIONSHIPS

2.4.1 ONTOLOGICAL REPRESENTATION OF SOCIAL INDIVIDUALS

(i) The Friend-of-a-Friend (FOAF) ontology that we use in our work is an OWL based format for representing personal information

(ii) FOAF started as experimentation with Semantic Web technology.

(iii) The idea of FOAF was to provide a machine processable format for representing the kind of information that made the original Web successful, namely the kind of personal information described in homepages of individuals.

(iv) Thus FOAF has a vocabulary for describing personal attribute information typically found on homepages such as name and email address of the individual, projects, interests, links to work and school homepage etc.

(v) FOAF profiles contain a description of friends the using the individuals same vocabulary that is used to describe the individual himself.

(vi) FOAF became the center point of interest in 2003 with the spread of SocialNetworking Services such Friendster, Orkut, LinkedIn etc.

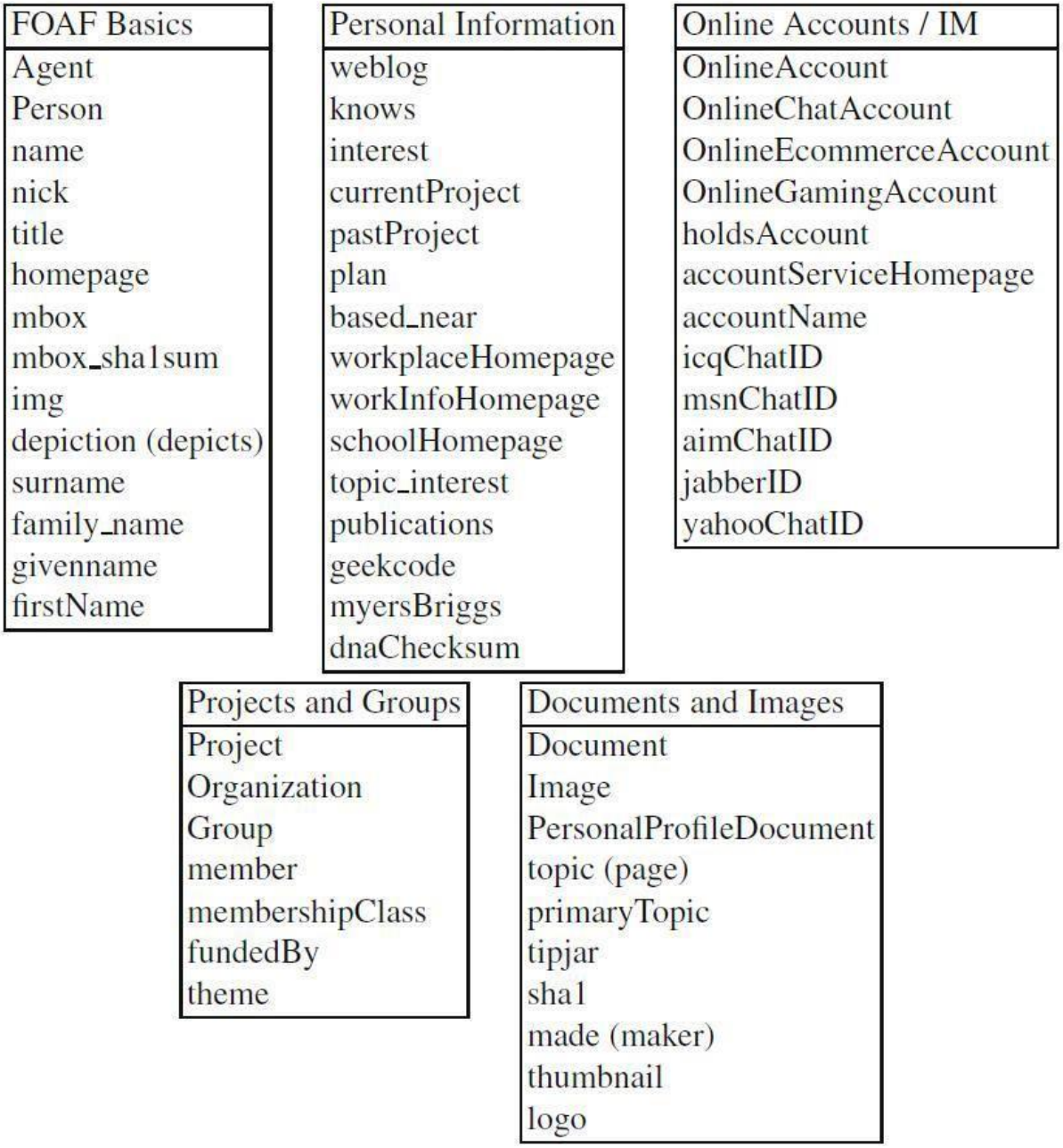
Drawbacks:

1. The information is under the control of the database owner
2. Centralized systems do not allow users to control the informationthey provide on their own terms.

(vii) FOAF profiles are created and controlled by the individual user and shared

in a distributed fashion. FOAF profiles are typically posted on the personal website of the user and linked from the home page user switch the HTML META tag.

(viii) An advantage of FOAF in terms of sharing FOAF data is the relative stability of the ontology. The number of FOAF users means that the maintainers of the ontology are obliged to keep the vocabulary and its semantics stable.



For example, the SIOC (Semantically Enabled Online Communities) project aims at connecting discussions across various types of for a Usenet, discussion boards, blogs, mailing lists etc by exposing the postings according to a shared ontology.

The key concepts of this ontology are the sioc:User account that is used to create a sioc:Post, which is part of a sioc:Forum at a certain sioc:Site. A sioc: User is not a subclass of foaf :Person (as a person may have multiple accounts), but related to the description of a person using the sioc:account of property. While FOAF has a rich ontology for characterizing individuals—especially with respect to their online presence—, but it is rather poor as a vocabulary for describing relationships.

2.4.2 ONTOLOGICAL REPRESENTATION OF SOCIAL RELATIONSHIPS

Ontological representations of social networks such as FOAF need to be extended with a framework for modeling and characterizing social relationships for two principle reasons:

(1) To support the automated integration of social information on a semantical basis and
(2) To capture established concepts in Social Network Analysis.

Characteristics of social relationships

- **Sign**: A relationship can represent both positive and negative attitudes such as likeor hate. The positive or negative charge of relationships is the subject of balance theory
- **Strength:** Tie strength itself is a complex construct of several characteristics of social relations. Tie strength lists the following: Frequency/frequent contact , Reciprocity, Trust/enforceable trust, Complementarity, Accommodation/adaptation, Indebtedness/imbalance, Collaboration, Transaction investments, Strong history, Fungible skills, Expectations, Social capital
- **Provenance**: A social relationship may be viewed differently by the individual participants of the relationship, sometimes even to the degree that the tie is unreciprocated. Similarly, outsiders may provide different accounts of the relationship, which is a well-known bias in SNA.
- **Relationship history**: Social relationships come into existence by some event involving two individuals
- **Relationship roles**: A social relationship may have a number of social roles associated with it, which we call relationship roles. For example, in a student/professor relationship within a university setting there is one individual playing the role of professor, while another individual is playing the role of a

student. Both the relationship and the roles may be limited in their interpretationand use to a certain social context.

Ideally, all users of all these services would agree to a single shared typology of social relations and shared characterizations of relations. However, this is neither feasible nor necessary. What is required from such a representation is that it is minimal in order to facilitate adoption and that it should preserve key identifying characteristics such as thecase of identifying properties for social individuals.

Conceptual model

Social relations could be represented as n-ary predicates; however, n-ary relations are not supported directly by the RDF/OWL languages. There are several alternatives to n- ary relations in RDF/ OWL

In all cases dealing with n-ary relations we employ the technique that is known as *reification*: we represent the relation as a class, whose instances are concrete relations of that type.

One may recall that RDF itself has a reified representation of statements: the *rdf :Statement* object represents the class of statements.

This class has three properties that correspond to the components of a statement, namely ***rdf: subject, rdf :predicate, rdf :object.***

These properties are used to link the statement instance to the resources involved in thestatement.

In other words relationships become subclasses of the ***rdf :Statement*** class. Common is that the new Relationship class is related to a general Parameter class by the **hasParameter** relationship. Relationship types such as Friendship are subclasses of the Relationship class, while their parameters (such as strength or frequency) are subtypes of the Parameter class.

Two alternatives:

The **first scheme** borrows from the design of OWL-S for representing service parameters, as used in the specification of the profile of a Web Service. Here, parameters are related by the valued-by metaproperty to their range. For example in an application Strength may be a subclass of Parameter valued-by integers. The **disadvantage** of this solution is that specifying values requires two statements or the introduction of a

constructed property.

 The **second alternative** differs in that the "native "representing meth parameters: the generic Parameter class is defined as a subclass of *rdf* :*Property*. This model has the advantage that it becomes more natural to represent parameter values and restrictions on them. The **disadvantage** is that this solution is not compliant with OWL.

DL Social relations are socially constructed objects: they are constructed in social environments by assigning a label to a common pattern of interaction between individuals.

Cognitive structuring, works by applying the generic pattern we associate with such a relationship to the actual state-of-affairs we observe. For example, a student/professor relationship at the Free University of Amsterdam is defined by the social context of the university and this kind of relationship may not be recognizable outside of the university.

The below figure shows descriptions and Situations ontology design pattern that providesa model of context and allows to clearly delineate these two layers of representation.

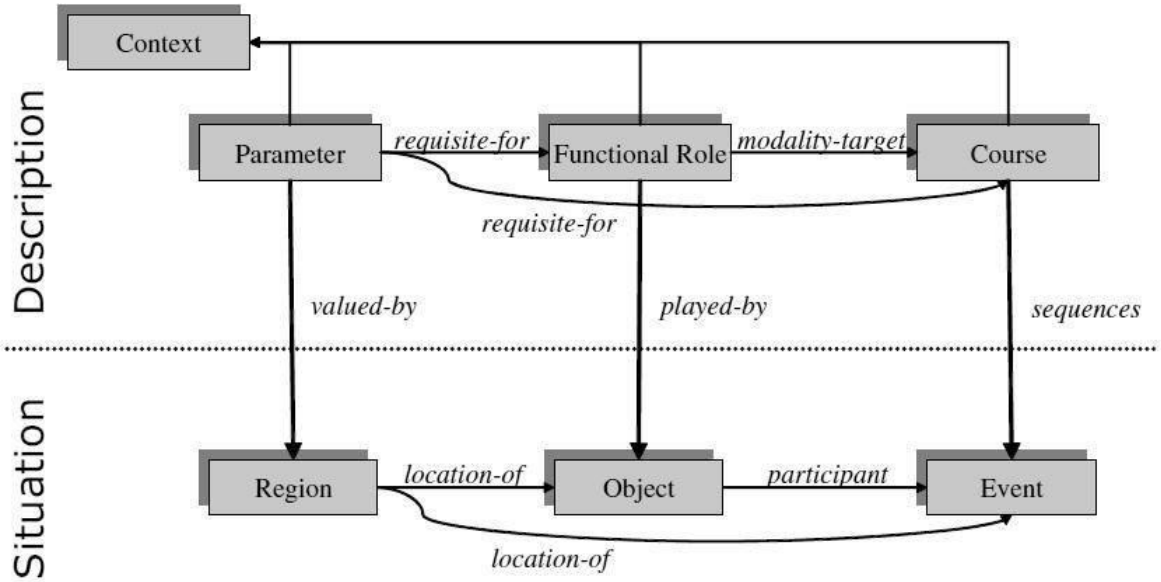


Fig. The Descriptions and Situations ontology design pattern

D&S is a generic pattern for modeling non-physical objects whose intended meaning results from statements, i.e. it emerges in combination with other entities. For example, a norm, a plan, or a social role is usually represented as a set of statements and not as a concept.

D&S is an ontology-design pattern in the sense that it is used as a template for creating domain ontologies in complex areas. D & S has been successfully applied in a wide range of real-life ontology engineering projects from representing Service Level Agreements (SLAs) to the descriptions of Web Services.

2.5 AGGREGATING AND REASONING WITH SOCIAL NETWORK DATA

2.5.1 ADVANCED REPRESENTATIONS

EXTRACTING EVOLUTION OF WEB COMMUNITY FROM A SERIES OF WEBARCHIVE

The extraction of Web community utilizes Web community chart A graph of communities, in which related communities are connected by weighted edges. The main advantage of the Web community chart is existence of relevance between communities.

2.5.1 **Notations Used**

- *t*1, *t*2, ..., *tn*: Time when each archive crawled. Currently, a month isused as the unit time
- *W*(*tk*): The Web archive at time *tk*
- *C*(*tk*): The Web community chart at time *tk*
- *c*(*tk*), *d*(*tk*), *e*(*tk*), ...: Communities in *C*(*tk*)

Types of Changes

- ✓ **Emerge** A community *c*(*tk*) emerges in *C*(*tk*), when *c*(*tk*) shares noURLs with any community in *C*(*tk*−1).

- ✓ **Dissolve**

A community *c*(*tk*−1) in *C*(*tk*) has dissolved, when *c*(*tk*−1) sharesno URLs with any community in *C*(*tk*)

✓ **Growth and Shrink**

The community grows when new URLs are appeared in *c*(*tk*), andshrinks when URLs disappeared from *c*(*tk*−1).

✓ **Split**

c(tk−1) shares URLs with multiple communities in *C*(*tk*)

✓ **Merge**

When multiple communities (*c*(*tk*−1)), *d*(*tk*−1), ...) share URLs with asingle community *e*(*tk*), these communities are merged into *e*(*tk*)

Evolution Metrics

Evolution metrics measure how a particular community c(tk) has evolved. The metricsare defined by differences between c(tk) and its corresponding community c(tk−1).

Growth Rate

The growth rate, Rgrow(c(tk−1), c(tk)), represents the increase of URLs per unit time. Itallows us to find most growing or shrinking communities.

$$R_{grow}(c(t_{k-1}), c(t_k)) = \frac{N(c(t_k)) - N(c(t_{k-1}))}{t_k - t_{k-1}}$$

Stability

Represents the amount of disappeared, appeared, merged and split URLs perunit time. A stable community on a topic is the best starting point for finding interesting changes around the topic.

$$R_{stability}(c(t_{k-1}),c(t_k)) = \frac{N(c(t_k)) + N(c(t_{k-1})) - 2N_{sh}(c(t_{k-1}),c(t_k))}{t_k - t_{k-1}}$$

Disappearance rate

The number of disappeared URLs from c(tk−1) per unit time. Higher disappear rate means that the community has lost URLs mainly by disappearance.

$$R_{disappear}(c(t_{k-1}),c(t_k)) = \frac{N_{dis}(c(t_{k-1}))}{t_k - t_{k-1}}$$

Merge rate

The number of absorbed URLs from other communities by merging per unit time. Higher merge rate means that the community has obtained URLs mainly by merging.

$$R_{merge}(c(t_{k-1}),c(t_k)) = \frac{N_{mg}(c(t_{k-1}))}{t_k - t_{k-1}}$$

Split Rate

The split rate, Rsplit (c(tk−1), c(tk)), is the number of split URLs from c(tk−1) per unit time. When the split rate is low, c(tk) is larger than other split communities. Otherwise, c(tk) is smaller than other split communities.

$$R_{split}(c(t_{k-1}),c(t_k)) = \frac{N_{sp}(c(t_{k-1}))}{t_k - t_{k-1}}$$

Other Metrics

The novelty metrics of a main line (*c*(*ti*), *c*(*ti*+1), ..., *c*(*t j*)) is calculated as follows.

$$R_{novelty}(c(t_i), c(t_j)) = \frac{\sum_{k=i}^{j} N_{ap}(c(t_k))}{t_j - t_i}$$

Web Archives and Graphs

- Web archiving is the process of collecting portions of the Web to ensure theinformation is preserved in an archive

- Web crawlers are used for automated capture due to the massive size andamount of information on the Web.

- From each archive, a Web graph is built with URLs and links by extractinganchors from all pages in the archive.

- The graph included not only URLs inside the archive, but also URLs outsidepointed to by inside URLs.

- By comparing these graphs, the Web was extremely dynamic

The size distribution of communities also follows the power law and its exponent did not change so much over time. Although the size distribution of communities is stable, the structure of communities changes dynamically. The structure of the chart changes mainly by split and merge, in which more than half of communities are involved.

Split and Merged Communities

- Both distributions roughly follow the power law, and show that split ormerge rate is small in most cases.

- Their shapes and scales are also similar.

- This symmetry is part of the reason why the size distribution ofcommunities does not change so much.

Emerged and Dissolved Communities

- The size distributions of emerged and dissolved communities alsofollow the power law
- Contribute to preserve the size distribution of communities.
- Small communities are easy to emerge and dissolve

Growth Rate

- The growth rate is small for most of communities, and the graph hasclear y-axis symmetry.
- Size distribution of communities is preserved over time.

Combining evolution metrics and relevance, evolution around a particular community can be located. The size distribution of communities followed the power-law, and its exponent did not change so much over time.

2.6 DETECTING COMMUNITIES IN SOCIAL NETWORKS

Detecting communities from given social networks are practically important for the following reasons:

1. Communities can be used for information recommendation because members of the communities often have similar tastes and preferences. Membership of detectedcommunities will be the basis of collaborative filtering.

2. Communities will help us understand the structures of given social networks. Communities are regarded as components of given social networks, and they will clarify the functions and properties of the networks.

3. Communities will play important roles when we visualize large-scale social networks. Relations of the communities clarify the processes of information sharing and information diffusions, and they may give us some insights for the growth the networks in the future.

2.7 EVALUATING COMMUNITIES

It is necessary to establish which partition exhibit a real community structure. Therefore, a quality function for evaluating how good a partition is needed. The most popular quality function is the modularity of Newman and Girivan:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

where the sum runs over all pairs of vertices, A is the adjacency matrix, ki is thedegree of vertex i and m is the total number of edges of the network. Modularitycan be rewritten as follows:

$$Q = \sum_{s=1}^{nm}\left[\frac{l_s}{m} - \left(\frac{d_s}{2m}\right)^2\right]$$

where nm is the number of communities, ls is the total number of edges joining vertices of community s, and ds is the sum of the degrees of the vertices of s. The first term of each summand is the fraction of edges of the network inside the community, whereas the second term represents the expected fraction of edges that

would be there if the network were a random network with the same degree foreach vertex. Figure 3.b illustrates the meaning of modularity.

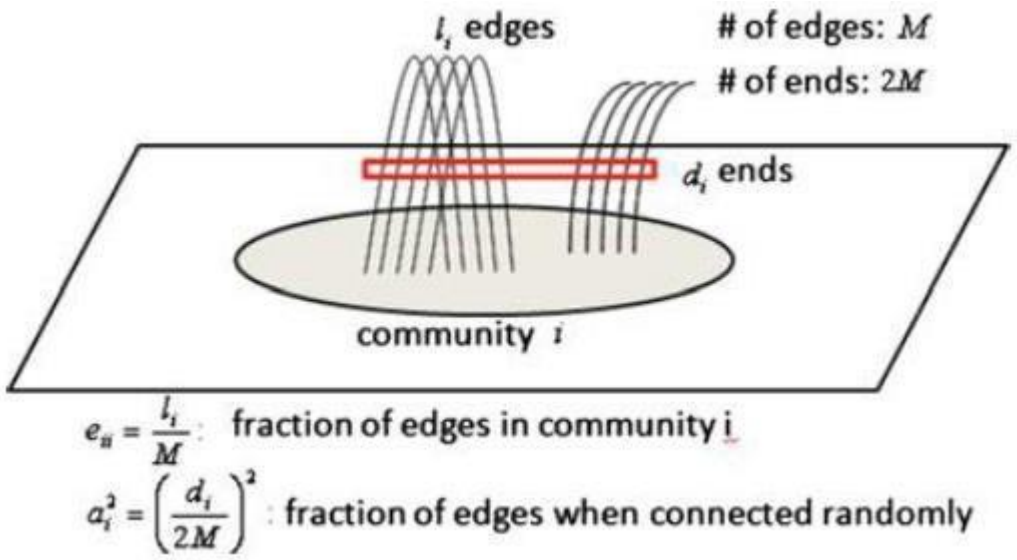


Fig.1 Modularity

The latter formula implicitly shows the definition of a community: a sub network isa community if the number of edges inside it is larger than the expected number in modularity's null model. The modularity of the whole network, taken as a single community, is zero. Modularity is always smaller than one, and it can be negative as well.

2.8 CORE METHODS FOR COMMUNITY DETECTION AND MINING

There are naive methods for dividing given networks into sub networks, such as graph partitioning, hierarchical clustering, and k-means clustering. The methods for detecting communities are roughly classified into the following categories:

(1) Divisive algorithms

(2) Modularity optimization

(3) Spectral algorithms and

(4) Other algorithms

Divisive Algorithms:

A simple way to identify communities in a network is to detect the edges that connect vertices of different communities and remove them, so that the communities get disconnected from each other. The steps of the algorithm are as follows:

(1) Computation of the centrality of all edges,

(2) Removal of edge with largest centrality,

(3) Recalculation of centralities on the running network, and

(4) Iteration of the cycle from step (2).

Edge betweenness is the number of shortest paths between all vertex pairs that run along the edge.

Modularity Optimization:

Modularity is a quality function for evaluating partitions. Therefore, the partition corresponding to its maximum value on a given network should be the best one. This is the main idea for modularity optimization. It has been proved that modularity optimization is an NPhard problem. However, 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. Another example of the algorithms aregreedy algorithms and simulated annealing.

Spectral Algorithms:

Spectral algorithms are to cut given network into pieces so that the number of edges to be cut will be minimized. One of the basic algorithms is spectral graph bipartitioning. The Laplacian matrix L of a network is an n * n symmetric matrix, with one row and column for each vertex. Laplacian matrix is defined as L =D - A

, where A is the adjacency matrix and D is the diagonal degree matrix with

$$Dii = \sum_k A_{ik}$$

All eigenvalues of L are real and non-negative, and L has a full set of n real and orthogonal eigenvectors. In order to minimize the above cut, vertices are partitioned based on the signs of the eigenvector that corresponds to the second smallest eigenvalue of L. In general, community detection based on repetitive bipartitioning is relatively fast.

Other Algorithms:

There are many other algorithms for detecting communities, such as the methods focusing on random walk, and the ones searching for overlapping cliques.

2.9 APPLICATIONS OF COMMUNITY MINING ALGORITHMS

Some applications of community mining, with respect to various tasks in social network analysis are listed below:

Network Reduction:

Network reduction is an important step in analyzing social networks. The example discussed here is taken from the work in which the network was

constructed from the bibliography of the book entitled “graph products: structure and recognition”. The bibliography contains 360 papers written by 314 authors.

Its corresponding network is a bipartite graph, in which each node denotes either one author or one paper, and link (i, j) represents author i publishing a paper
j. Community structure is detected using a community mining algorithm called ICS. Each community contains some papers and their corresponding coauthors.

Most of the detected communities are self-connected components. Moreover, the clustered coauthor network can be reduced into a much smaller one by condensing each community as one node. Finally, the top-level condensed network corresponding to a 3-community structure is constructed by using ICS from the condensed network. From this a dendrogram corresponding to the originalcoauthor network can be built.

Discovering Scientific Collaboration Groups from Social Networks

This section show how community mining techniques can be applied to the analysis of scientific collaborations among researchers. Flink is a social network that describes the scientific collaborations among 681 semantic Web researchers (http://flink.semanticweb.org/).

The network was constructed based on semantic Web technologies and all related semantic information was automatically extracted from “Web-accessible information sources”, such as “Web pages, FOAF profiles, email lists, and publication archives”. The weights on the links measure the degrees of collaboration.

Mining Communities from Distributed and Dynamic Networks:

Many applications involve distributed and dynamically-evolving networks, in which resources and controls are not only decentralized but also updated frequently. One promising solution is based on an Autonomy-Oriented Computing (AOC) approach, in which a group of self-organizing agents are utilized. The agents will rely only on their locally acquired information about networks. Intelligent Portable Digital Assistants (or iPDAs for short) that people carry aroundcan form a distributed network, in which their users communicate with each other through calls or messages.

One useful function of iPDAs would be to find and recommend new friends with common interests, or potential partners in research or business, to the users. The way to implement it will be through the following steps:

(1) Based on an iPDA user's communication traces, selecting individuals who havefrequently contacted or been contacted with the user during a certain period of time;

(2) Taking the selected individuals as the input to an AOC-based algorithm.

(3) Ranking and recommending new persons who might not be included the current acquaintance book, the user.

In such a way, people can periodically receive recommendations about friends orpartners from their iPDAs.

SOCIAL NETWORK SOFTWARE

SNA software generates features from raw network data formatted in an edgelist, adjacency list, or adjacency matrix (also called sociomatrix), often combined with (individual/node-level) attribute data. Though the majority of network analysis software uses a plain text ASCII data format, some software packages contain the capability to utilize relational databases to import and/or store network features. visual representations of social networks are important to

understand network data and convey the result of the analysis. Visualization often also facilitates qualitative interpretation of network data. With respect to visualization, network analysis tools are used to change the layout, colors, size and other properties of the network representation.

Some SNA software can perform predictive analysis.This includes using network phenomena such as a tie to predict individual level outcomes (often called peer influence or contagion modeling), using individual-level phenomena to predict network outcomes such as the formation of a tie/edge (often called homophily models) or particular type of triad, or using network phenomena to predict other network phenomena, such as using a triad formation at time 0 to predict tie formation at time 1.

PRODUCT	**MAIN FUNCTIONALITY**	**PLATFORM**	**LICEN AND SE COST**	**NOTES**
Allegrograph	Graph Database. RDF with Gruff visualization tool	Linux, Mac, Windows	Free and Commercia –	AllegroGraphis a graph database. It is disk-based, full y transactional OLT P database that stores data structured in graphs rather than in tables. AllegroGra ph includes a Social Networking Analyti cs library.
Gephi	Graph exploration and manipulation software	Any system supporting Java 1.6 and OpenGL	Open Source (GPL 3) seeking contributors	Gephi[1] is an interactive visualization an d exploration platform for all kinds of networks and complex system s, dynamic and hierarchical graphs. It is a tool for people that have to

				explore and understand graphs. The user interacts with the representation.
Graph Stream	Dynamic Graph Library	Any system supporting Java	Open Source	With GraphStream you deal with graphs. Static and Dynamic. You create them from scratch, from a file or any source. You display and render them.

Java Universal Network/Graph (JUNG) Framework	network and graph manipulation, analysis, and visualization	Any platformsupporting Java	Open source (BSDlicense)	JUNG is a Java API and library that provides a common and extensible language for the modeling, analysis, and visualization of relational data. It supports a variety of graph types (including hypergraphs), supports graph elements of any type and with any properties
Mathematica	Graph analysis,statistics, data visualization, optimization, image recognitio	Windows, Macintosh, Linux	Commercial	Mathematica is a general purpose computation and analysis environment.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT – III – SOCIAL NETWORK ANALYSIS – SITA3005

UNIT 3 CLIQUES, CLUSTERS AND COMPONENTS 9 Hrs.

Components and Subgraphs: Sub graphs - Ego Networks, Triads, Cliques, Hierarchical Clustering, Triads, Network Density and conflict. Density: Egocentric and Sociocentric - Digression on Absolute Density – Community structure and Density, Centrality : Local and Global - Centralization and Graph Centres, Cliques and their intersections, Components and Citation Circles - Positions, Sets and Clusters.

Components and Subgraphs

A subgraph is a subset of the nodes of a network, and all of the edges linking these nodes. Any group of nodes can form a subgraph—and further down we will describe several interesting ways to use this.

Component subgraphs (or simply components) are portions of the network that are disconnected from each other. Before the meeting of Romeo and Juliet, the two families were quite separate (save for the conflict ties), and thus could be treated as components.

Many real networks (especially these collected with random sampling) have multiple components. One could argue that this is a sampling error (which is very possible)—but at the same time, it may just mean that the ties between components are outside of the scope of the sampling and may in fact be irrelevant.

Blocks and Cutpoints (Bi-components)

An alternative approach to finding the key "weak" spots in the graph is to ask: if a node were removed, would the structure become divided into un-connected parts? If there are such nodes, they are called "cutpoints." And, one can imagine that such cutpoints may be particularly important actors -- who may act as brokers among otherwise disconnected groups. The divisions into which cut- points divide a graph are called blocks. We can find the maximal non-separable sub-graphs (blocks) of a graph by locating the cutpoints. That is, we try to find the nodes that connects the graph (if there are any). Another name for a block isa "bi-component."

Factions

Imagine a society in which each person was closely tied to all others in their ownsub-population (that is, all sub-populations are cliques), and there are no connections at all among sub-populations (that is, each sub-population is a component). Most real populations do not look like this, but the "ideal type" of complete connection within and complete disconnection between sub-groups is a useful reference point for assessing the degree of "factionalization" in a population.If we took all the members of each "faction" in this ideal-typical society, and put their rows and columns together in an adjacency matrix (i.e. permuted the matrix), we would see a distinctive pattern of "1-blocks" and "0- blocks." All connections among actors within a faction would be present, all connections between actors in different factions would be absent.

Network Measurements

A number of measurable network characteristics were developed to gain a greater insight into networks, with many of them having their roots in social studies on the relationships among social actors. In this section, we will discuss three categories of measurements that have been defined in the social network analysis stream:

1. Network connection, which includes transitivity, multiplexity, homophily, dyads and mutuality, balance and triads, and reciprocity
2. Network distribution, which includes the distance between nodes, degree cen- trality, closeness centrality, betweenness centrality, eigenvector centrality and density

3. Network segmentation, which includes cohesive subgroups, cliques, clusteringcoefficient, k-cores, core/periphery, block models, and hierarchical clustering

Network Connection

Network connection (or connectivity) refers to the ability to move from one nodeto another in a network. It is the ratio between route distance and geodesic distance. Connectivity can be calculated locally (for a part of the network) and globally (for the entire network). Let's take a look at some of the important metrics of network connection.

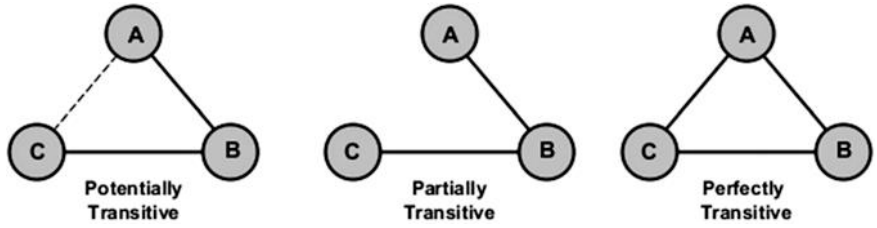


Fig.3.1 Transitivity between nodes

Transitivity

Transitivity is a network property that refers to the extent to which a relation between two nodes is transitive. It is a very important measure in social networksbut less important in other types of networks. In social networks, the term transitivity reflects the friend-of-a-friend concept. It is sometimes used as a synonym of whole-network clustering coefficient.

Homophily

Homophily is the tendency of individuals to connect with others who share the same attitudes and beliefs. The tendency of individuals to associate with similar others based on gender, education, race, or other socioeconomic characteristics isvery common in social communities. Coordination and cooperation are typically more successful between people who show some similarity to each other such that individuals in homophilic relationships are likely to hear about new ideas or ask for help from each other.Homophily in the context of online social networking can be understood from the similarity of users who are using the network in terms of age, educational back- ground, region, or profession. In the sense of corporate networks, homophily is translated as the similarity of professional or academic qualifications.

Balance and Triads

A triad is a network structure consisting of three actors and three dyads. Given a complete graph of three actors (triad), we can identify four different types of rela- tionships, depending on the number of negative relationships between nodes: (a) a friend of my friend is my friend, (b) an enemy of my enemy is my friend, (c) a friend of my friend is my enemy, and (d) an enemy of my enemy is my enemy.Triads with an odd number of "+" edges are bal- anced, while triads with an even number of "-" edges are unbalanced. Imbalanced graph configurations usually create stress for individuals located on them.

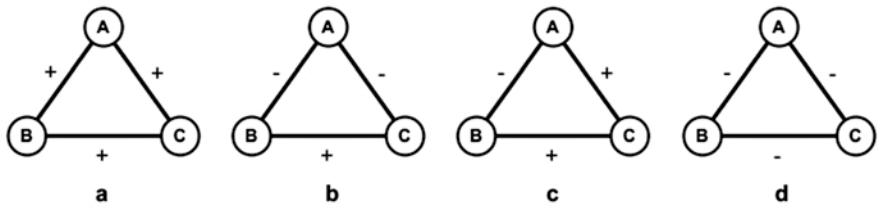


Fig. 3.2 Graph with three nodes in four states

The above graph is a type of signed graphs which have been studied since the 1950s. They are a special case of valued graphs in which ties are allowed to haveone of two opposing values to convey the positive or negative sentiment. Examples of signed graphs include friend/foe, trust/distrust or like/dislike, esteem/disesteem, praise/blame, influence/negative influence, etc. They are very common in sociology and psychology but less common in fields such as physics and chemistry.

- In figure a, all the three actors have positive feelings, and there is no place for conflict among them. The configuration is coherent and lacks inner tensions between members.
- Figure b is also stable since two actors (B and C) share the same negative feeling towards actor A, but they like each other.
- Figure c is unstable because actors A and B have a negative feeling towards each other, while both have a positive feeling towards actor C which has to divide its loyalty between the other two actors.
- Figure d is also unstable and will eventually break down, as it has an odd number of negative signs.

In Fig. 3.2, b, types of balanced subgraphs are shown, whereas in Fig. 3.2 c, d, types of unbalanced graphs are presented. An obvious way to avoid unbalancesin subgraphs is by sign shifting, which includes changing signs such that enmities (negative signs) become friendships (positive signs) or vice versa. Within real net- works, stable configurations appear far more often than unstable configurations.It should be noted here that a negative sign between two nodes does not mean the lack of tie between these two nodes. While a negative sign between two nodes is a clear mark of an inimical relationship, the absence of a tiebetween these nodes suggests the absence of interaction or communication between them.

Reciprocity

Reciprocity is a measure of the tendency towards building mutually directed con-nections between two actors. It refers to the number of reciprocated tie for a specific actor in a network. For example, if *u* connects to *v*, then *v* connects to *u* and vice versa. In real life scenarios, it is important to know whether received help is also given or whether given help is translated as help by the receiver.For agiven node *v*, reciprocity is the ratio between the number of nodes which have both incoming and outgoing connections from/to *v*, to the number of nodes which only have incoming connections from *v*. For an entire network, reciprocityis calculated as the fraction of edges that are reciprocated. Average reciprocity is

calculated by averaging reciprocity values of all nodes in the network.

Network Distribution

Measurements of network distribution are related to how nodes and edges are distributed in a network.

Distance Between Two Nodes

Distance is a network metric that allows the calculation of the number of edges between any pair of nodes in a network. Measuring distances between nodes in graphs is critical for many implementations like graph clustering and outlier detection. Sometimes, the distance measure is used to see if the two nodes are similar or not. Any commonly used shortest path calculation algorithm (e.g., Dijkstra) can be used to provide all shortest paths in a network with their lengths.We can use the distance measure to calculate node eccentricity, which is the maximum distances from a given node to all other nodes in a network. It is also possible to calculate network diameter, which is the highest eccentricity of its nodes and thus represents the maximum distance between nodes.In most social networks, the shortest path is computed based on the cost of tran- sition from one node to another such that the longer the path value, the greater the cost.Within a community, there might be many edges between nodes, but between communities, there are fewer edges.

Degree Centrality

In degree centrality metric, the importance of a node is determined by how manynodes it is connected to. It is a measurement of the number of direct links to other actors in the network. This means that the larger the number of adjacent nodes, the more important the node since it is independent of other actors that reach great parts of the network. It is a local measure since its value is computed based on the number of links an actor has to the other actors directly adjacent to it. Actors in social net- works with a high degree of centrality serve as hubs and as major channels of information.In social networks, for example, node degree distribution follows a power law distribution, which means that very few nodes have an extremely large number of connections. Naturally, those high-degree nodes have more impact in the network than other nodes and thus are considered more important. A node i's degree central- ity $d(i)$ can be formulated as

$$d(i) = \sum_{j} m_{ij}$$

where m_{ij} = 1 if there is a link between nodes *i* and *j* and m_{ij} = 0 if there is no such link. For directed networks, it is important to differentiate between the in- degree centrality and the out-degree centrality.

Identifying individuals with the highest-degree centrality is essential in network analysis because having many ties means having multiple ways to fulfillthe require- ments of satisfying needs, becoming less dependent on other individuals, and hav- ing better access to network resources. Persons with the highest-degree centrality are often third parties and deal makers and able to benefit from this brokerage. For directed networks, in-degree is often used as a proxy for popularity .The figure shows that node A and node B are at exceptionalstructural positions. All communications lines must go through them. This gives us a conclu- sion that both nodes, A and B, are powerful merely because of their excellent posi- tions. However, such a finding is largely based on the nature of links and the nature of embedded relationships.

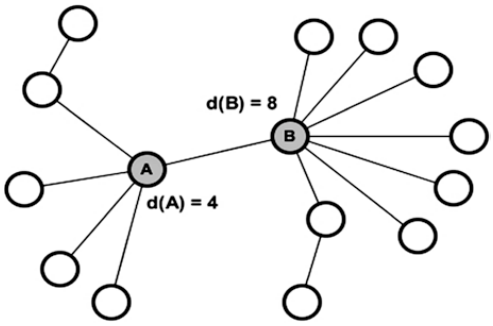


Fig. 3.3 Degree centrality of nodes

Closeness Centrality

Closeness centrality can be defined as how close, to a particular actor, other actors are. It is the sum of the geodesic distances of a node to all other nodes in the net- work. It computes the length of paths from one actor to other actors in the network.

That actor can be important if it is relatively close to the remaining set of actors in the network. The mathematical representation of closeness centrality, $C(i)$, is given as follows:

$$C(i) = \sum_j d_{ij}$$

where d_{ij} is the geodesic distance from node *i* to node *j* (number of links in the shortest path from node *i* to node *j*).

Closeness centrality is important to understand information dissemination in net- works in the way that the distance between one particular node and others has an effect on how this node can receive from or send information (e.g., gossip) to other nodes. In social networks, this ability is limited by what is called "horizon of observ- ability" which states that individuals have almost no sight into what is going on after two steps.Because closeness centrality is based on the distance between network nodes, it can be considered the inverse of centrality because large values refer to lower cen- trality, whereas small values refer to high centrality. Computationally, the value of *C(i)* is a number between 0 and 1, where higher numbers mean greater closeness (lower average distance) whereas lower numbers mean insignificant closeness (higher average distance) .In the figure, the nodes in gray are the most central regarding closeness because they can reach the rest of nodes in the network easily and equally.

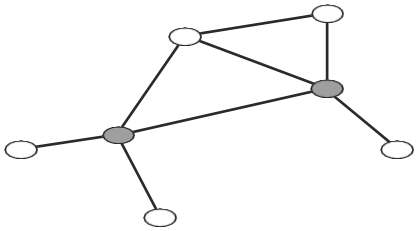


Fig3.4 Closeness centrality of nodes

They have the ability to reach all other nodes in the fastest amount of time. The other nodes lack these privileged positions.Because closeness centrality is based on shortest path calculations, its usefulness when applied to large networks can be brought into question in the way that closeness produces little variation in the results, which makes differentiating between nodes more difficult.In information networks, closeness reveals how long it takes for a bit of information to flow fromone node to others in the network. High-scoring nodes usually have shorter pathsto the rest of nodes in the network.

Betweenness Centrality

Betweenness centrality can be described as how important an actor is, as a link between different networks. It represents the number of times an actor needs to pass via a given actor to reach another actor. Nodes with high betweenness centrality control the flow of information because they form critical bridges between other actors or groups of actors. Betweenness centrality of node *i* is calculated as follows:

$$b(i) = \sum_{j,k} \frac{g_{jik}}{g_{jk}}$$

where g_{jk} is the number of shortest paths from node (*j*) to node *k* (*j* and $k \neq i$) and g_{jik} is the number of shortest paths from node (*j*) to node *k* passing through the node (*i*).

Eigenvector Centrality

Eigenvector centrality measurement describes the centrality of a person with regard to the global structure of the network. It assigns relative scores to all nodes in the network based on the concept that connections to nodes with high scoring contribute more to the score of the node in question than co*n*nections to nodes with low scoring.It measures the extent to which a node is connected to well-connected nodes. It is computed by taking the principal eigenvector of the adjacency matrix. Calculating centrality in the way that eigenvector measure proposes differs from the way that degree measure applies to calculate centrality which is based on simply adding up the number of links of each node.

Density

Density is defined as the degree to which network nodes are connected one to another. It can be used as a measure of how close a network is to complete. In the case of a complete graph (a graph in which all possible edges are present), density is equal to one. In real life, a dense group of objects has many connections among its entities (i.e., has a high density), while a sparse group has few of them (i.e., has a low density).

Formally, the density *D*(*G*) of graph *G* is defined as the fraction of edges in *G* to the number of all possible edges. Density values range between zero and one

[0, 1].

Proportion of ties between alters comparedto number possibleties

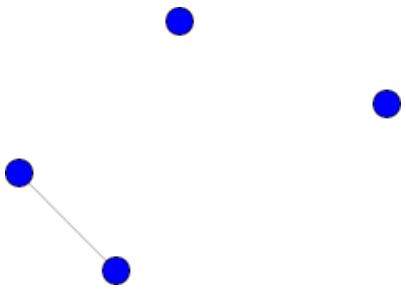


Fig.3.5 Ties

Total ties=1

No:of possible ties=6 (N*(N-1)/2)//N is the number of nodesDensity=1/6

Cohesive Subgroups

Cohesive groups are communities in which the nodes (members) are connectedto others in the same group more frequent than they are to those who are outside of the group, allowing all of the members of the group to reach each other. Within such a highly cohesive group, members tend to have strong homogenous beliefs. Connections between community members can be formed either through personal contacts (i.e., direct) or joint group membership (i.e., indirect). As such,the more tightly the individuals are tied into a community, the more they are affected by group standards.

Cliques

A clique is a graph (or subgraph) in which every node is connected to every othernode. Socially translated, a clique is a social grouping in which all individuals know each other (i.e., there is an edge between each pair of nodes). A triangle is an exam- ple of a clique of size three since it has three nodes and all the nodes are connected.A maximal clique is a clique that is not a subset of any other cliquein the graph. A clique with size greater than or equal to that of every other clique in the graph is called a maximum clique.

Relaxation of Strict Cliques

- Distance (length of paths)

- N-clique, n-clan, n-club

- Density (number of ties)
  - K-plex, ls-set, lambda set, k-core, component

N-cliques

The strict clique definition (maximal fully-connected sub-graph) may be too strong for many purposes. It insists that every member or a sub-group have a direct tie with each and every other member. You can probably think of cases of "cliques" where at least some members are not so tightly or closely connected. There are two major ways that the "clique" definition has been "relaxed" to try tomake it more helpful and general.

One alternative is to define an actor as a member of a clique if they are connected to every other member of the group at a distance greater than one. Usually, the path distance two is used. This corresponds to being "a friend of a friend." This approach to defining sub-structures is called N-clique, where N stands for the length of the path allowed to make a connection to all other members.

N-Clans

The N-clique approach tends to find long and stringy groupings rather than the tight and discrete ones of the maximal approach. In some cases, N-cliques can befound that have a property that is probably undesirable for many purposes: it is possible for members of N-cliques to be connected by actors who are not, themselves, members of the clique. For most sociological applications, this is quite troublesome.

To overcome this problem, some analysts have suggested restricting N-cliquesby insisting that the total span or path distance between any two members of an N-clique also satisfy a condition. The additional restriction has the effect of forcing all ties among members of an n-clique to occur by way of other membersof the n-clique.

Ego Net

The ego-net approach to social network analysis, which takes discrete individual actors and their contacts as its starting point, is one of the most widelyused approaches.

- Ego network (personal network)
- Ego: focal node/respondent
- Alter: actors ego has ties with
- Ties between alters

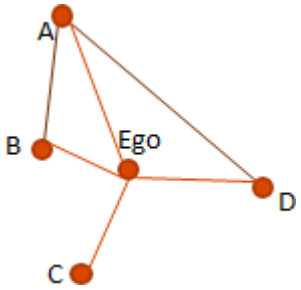


Fig 3.6 Ego network

Why use ego network data?

- From ego's perspective, personal network is important for:
  - Social support
  - Access to resources
  - Influence/normative pressure
- From a more global perspective, ego network data are useful for:
  - Studying mixing patterns between groups
  - Potential for diffusion
    - Disease propagation
    - Adoption of innovation: new product or health practice
- Lots can be had from ego network data!
  - Composition of individual's local social world
    - Demographic characteristics of alters
    - Shared health behaviors
  - Structural features
    - Size
    - Density
  - Nature of the ties
    - Frequency, duration, closeness
    - Specific exchanges

When to use Ego Network Analysis

- If your research question is about phenomena of or affecting individual entities across different settings (networks) use the ego- centric approach
  - Individual people, organizations, nations, etc.

- If your research question is about different patterns of interaction within defined groups (networks), use the socio- centric approach

- E.g., who are the key players in a group? How do ideas diffusethrough a group?

Which Theories are Ego-centric?

- **Most theories under the rubric of social capital areego-centric**
  - Topological
    - Structural holes / Brokerage
    - Embeddedness
- **Compositional**
    - Size
    - Alter attributes

Steps to a SNA study

1. **Identify the population**
   - Sampling, gaining access
2. **Determine the data sources**
   - Surveys, interviews, observations, archival
3. **Collect the data**
   - Instrument design

Step 1.Identify the Population

Sampling Criteria

- Determined by research question
  - High tech entrepreneurs
  - Alumni of defunct organizations
  - Basketball coaches
  - First time mothers returning to the workforce
  - Baseball Hall of Fame inductees
  - Contingent workers
  - People with invisible stigmatized identities

Step 1.Identify the Population

Gaining Access

- Same concerns as other research
  - It depends on the sensitivity of the questions that you areasking
    - Length of interview can be daunting
      - Depends on the number of alters

Step 2: Determine Data Sources

- Surveys
- Interviews
- Observations
- Archival data

Step 3: Collect the Data

- What data should you collect?
  - What questions need to be answered?
- How to format your data collection instrument (e.g., a survey,spreadsheet, database, etc.)?

Data Collection in an Ego-centric Study

1. Attributes about Ego
2. Name generator
   - Obtain a list of alters
3. Name interpreter
   - Assess ego's relationships with generated list of alters?
4. Alter Attributes
   - Collect data on the list of alters
5. Alter – Alter Relationships
   - Determine whether the listed alters are connected

Attributes about Ego

- Typical variables for case based analysis
  - Age
  - Gender
  - Education
  - Profession
  - SES
  - Etc.

Sample Name Generators

Questions that will elicit the names of alters

- From time to time, most people discuss important personal matters with other people. Looking back over the last six months who are the people with whom you discussed an important personal matter? Please just telI me their first namesor initials.

Consider the people with whom you like to spend your free time.

- Over the last six months, who are the one or two people you have been with the most often for informal social activities such as going out to lunch, dinner, drinks, films, visiting one another's homes, and so on?

Sample Name Interpreter

- Questions that deal with ego's relationshipwith [or perceptionof] each alter
  - How close are you with <alter>?
  - How frequently do you interact with <alter>?
  - How long have you known <alter>?
- All of these questions will be asked for eachalter named in theprevious section

Sample Alter Attribute Questions As far as you know, what is <alter>' s highest

- As far as you know, what is <alter>' s highest level of education?
  - Age, occupation, race, gender, nationality, salary, drug usehabits, etc
- Some approaches do not distinguish betweenname interpreters and alter attribute

Sample Alter-Alter Relationship Questions

- Think about the relationship between <alter1> and <alter2>. Would you say that they are strangers, just friends, or especially close?
- Note: this question is asked for each unique alter- alter pair. E.g., if there are 20 alters, there are 190 alter-alter relationship questions!
  - Typically, we only ask one alter-alter relationshipquestion

Why Ego-Centric Analysis

- **Asks different questions than whole networkanalysis.**
  - In fact, many of the various approaches to "Social Capital" lend themselves particularly to the analysis of Ego-Centric or Personal networks

Kinds of Analyses

- In Ego-Centric Network analyses we are typically looking to use network-derived measures as variables in more traditionalcase-based analyses
  - E.g., instead of just age, education, and family SES to predict earning potential, we might also include heterogeneity of network or brokerage statistics

Many different kinds of network measures, thesimplest is degree (size Data Analysis of Ego Networks

1. Size

- How many contacts does Ego have?

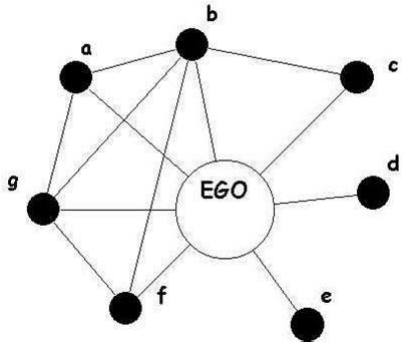
2. Composition

- What types of resources does ego have access to? (e.g., quality)
- Does ego interact with others like him/herself? (e.g., homophily)
- Are ego's alters all alike? (e.g., homogeneity?)

3. Structure

- Does ego connect otherwise unconnected alters? (e.g., brokerage,density, etc)
- Does ego have ties with non-redundant alters (e.g., effective size,efficiency, constraint)

Size



Degree = 7

Composition: Content

- The attributes (resources) of others to whom I am connectedaffect my success or opportunities
  - Access to resources or information

– Probability of exposure to/experience with

Composition: Similarity Between Ego & Alter

- **Homophily**
  - We may posit that a relationship exists between some phenomenon and whether or not ego and alters in a network share an attribute
    - Selection
      - Teens who smoke tend to choose friends who also smoke
    - Influence
      - Overtime, having a network dominated by people with particular views may lead to one taking on those views

Composition: Homophily

- A CFO who surrounds herself with all finance people
- A Politician who surrounds himself with all members of thesame political party

Composition: Dissimilarity Between Ego & Alter

Heterophily

– We may posit that a relationship exists between some phenomenon and adifference between ego and alters along some attribute

- Mentoring tends to be heterophilous with age

Composition: Homophily/Heterophily
Krackhardt and Stern's E-I index

$$\frac{E-I}{E+I}$$

• E is number of ties to members in different groups (external), I is number of ties to members of same group (internal).Varies between -1 (homophily) and +1 (heterophily)

Composition: Heterogeneity

- Similar to homophily, but distinct in that it looks not atsimilarityto ego, but just among the alters

- Diversity on some attribute may be provide access to differentinformation, opinions, opportunities, etc.

  - My views about social welfare may be affected by the diversity in SES present in my personal network (irrespective of or in addition to my own SES)

Structural Analyses

- Burt's work is particularly and explicitly ego-network based in calculation

  - My opportunities are affected by the connections that exist (or are absent) between those to whomI am connected

Structural Holes

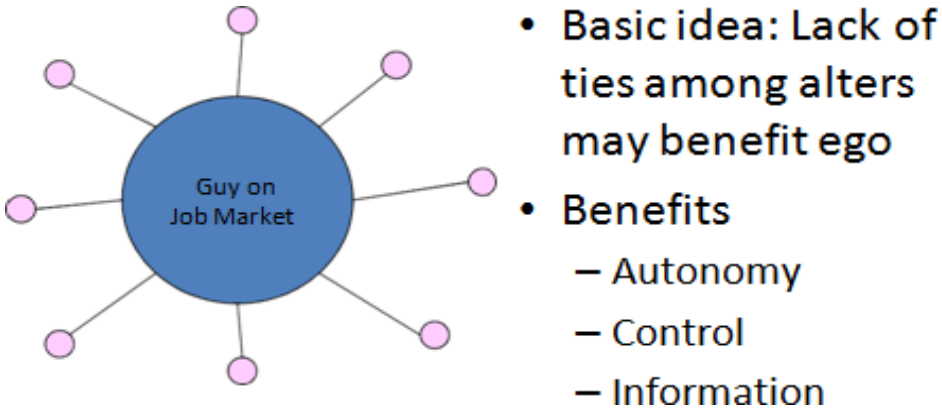


Fig 3.7 Structural holes

Burt's Measures of Structural Holes

- Effective size
- Efficiency
- Constraint

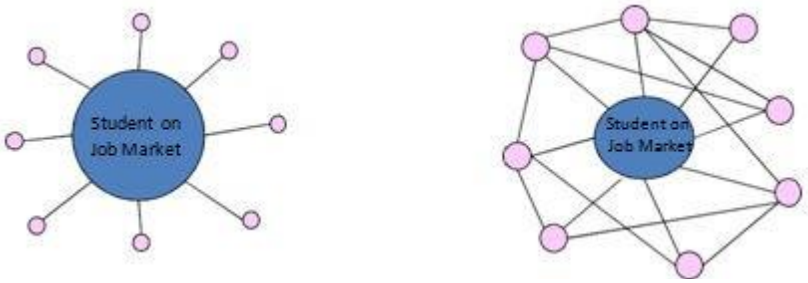
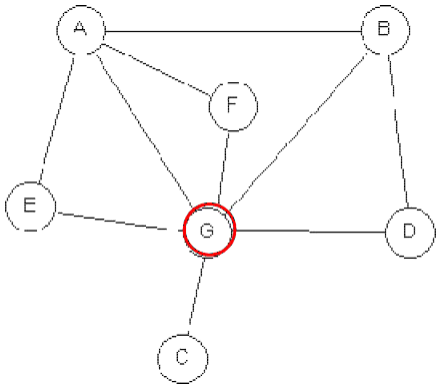


Fig 3.8 Structural holes

Effective Size



Node "G" is EGO	A	B	C	D	E	F	Total
Redundancy with EGO's other Alters:	3/6	2/6	0/6	1/6	1/6	1/6	1.33

Fig 3.9 Effective Size

Effective Size of G = Number of G's Alters – Sum of Redundancy of G'salters

= 6 – 1.33 = 4.67

Efficiency

Efficiency = (Effective Size) / (Actual Size)

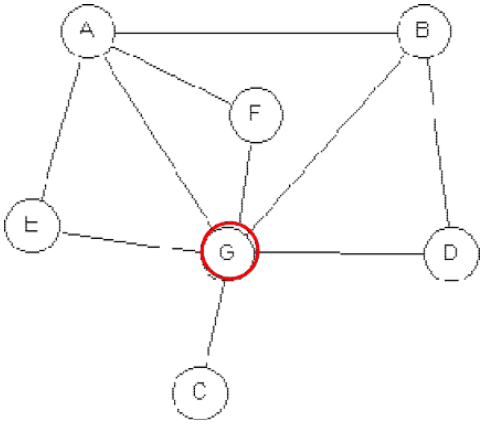


Fig 3.10 Efficiency

Actual Size = 6

Effective Size of G = 4.67

Efficiency = 4.67/6 = ~0.78

Constraint: The Basic Idea

- Constraint is a summary measure that taps the extent to whichego's connections are to others whoare connected to one another.

- If ego's boyfriend bowls with her brother and father every Wednesday night, she may be constrained in terms of distancing herself from him, even if they break up.

- There's a normative bias in much of the literature that lessconstraint is good

Ego-Centric Network Analysis

- When conducted across many, independent egos, presentsdifferent problems

- Many Social Network Analysis tools ill suited to the natureof such analyses
  - Really designed for "whole network" analysis

- Ego Network analyses require either:
  - joining into one large, sparse, blocked network, or
  - repetition of analysis of individual networks

- Can be tedious if there's no facility for batching them

Local and Global Measures

Local: personal network size

- Number of alters (social support) predicting health outcomes
- Number of drug partners predicting future risky behaviorGlobal: degree distribution by aggregating over all cases
- Distribution of ties per person

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT – IV – SOCIAL NETWORK ANALYSIS – SITA3005

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.

4.1 Understanding and predicting human behavior for social communities

With the rapid advance in technology, it is becoming increasingly feasible for people to take advantage of the devices and services in the surrounding environment to remain "connected" and continuously enjoy the activity they are engaged in, be it sports, entertainment, or work.

Ubiquitous computing environment will allow everyone permanent access to the Internet anytime, anywhere and anyhow.

QoE (Quality of Experience) is a consequence of a user's internal state (e.g., predispositions, expectations, needs, motivation, mood), the characteristics of the designed system (e.g., usability, functionality, relevance) and the context (or the environment) within which the interaction occurs (e.g., social setting, meaningfulness of the activity).

4.2 User Data Management, Inference and Distribution

- User-oriented creation/execution environments lack on the capability to adapt to the heterogeneity of devices, technologies and the specificity of each individualuser.
- For user flexibility and personalization requires user profile management systems which include limited information about user preferences and contexts.
- 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.
- These efforts have been taken by European Telecommunications Standards Institute (ETSI), the Third Generation Partnership Project (3GPP), Open Mobile Alliance (OMA)
- Considering data requirements from a wide range of facilities the concept of Common Profile Storage (CPS) is defined by 3GPP.

- as a framework for streamlining service-independent user data and storing it under a single logical structure in order to avoid duplications and data inconsistency

- Logically centralized data storage can be mapped to physically distributed configurations and should allow data to be accessed in a standard format
- Data storage can be grouped into three main classes: the syntactic, semantic and modeling approaches which enable interoperability of user profile data management for a Future Internet.
- To improve the degree of services personalization it is important to generate new information from the existing one.
- In this sense, social networks, user modeling and reality mining techniques can be empowered to study patterns and predict future behaviors.
- The basic motivation is the demand to exploit knowledge from various amounts of data collected, pertaining to social behavior of users in online environments.
- To handle complex situations, the concept of decomposition is applied to the situation into a hierarchy of sub-situations.
- These sub-situations can be handled autonomously with respect to sensing and reasoning. The handling of complex situations can be simplified by decomposition.
- Another similar perspective is called **layered reasoning**, where
  - the first stage involves feature extraction and grouping (i.e., resulting in low-level context),
  - the second event, state and activity recognition (i.e., originating mid-level context), while
  - the last stage is dedicated to prediction and inference of new knowledge
- Research in Social network usually focuses on
  - quantifying or qualifying the relationship between peers,
  - where algorithms such as centrality and prestige can be used to calculate the proximity, influence or importance of a node in a network,
  - while clustering and classification can be applied to similarity computation, respectively.

Combining all of pre-enunciated concepts with ontologies and semantic technologies, we present a generic framework for managing user related data, will pave the way to understanding and predicting future human behavior within social communities.

4.3 Enabling New Human Experiences

It is important to understand what are the technologies behind user data management, how to link them and what can they achieve when combined in synergy.

4.3.1. The Technologies

a. Social Networks

- Humans in all cultures at all times form complex social networks
- Social network - means ongoing relations among people that matter to those engaged in the group, either for specific reasons or for more general expressions of mutual agreement.
- Social networks among individuals who may not be related can be validated and maintained by agreement on objectives, social values, or even by choice of entertainment.
    - involve reciprocal responsibilities and roles that may be selfless or self-interest based.
- Social networks are trusted because of shared experiences and the perception of shared values and shared needs
- Behavior of individuals in online networks can be slightly different from the same individuals interacting in a more traditional social network (reality).
- It gives us invaluable approaches on the people we are communicating with, which groups are we engaged, which are our preferences, etc.

b. Reality Mining

- To overcome the differences between online and "offline" networks, reality mining techniques can be empowered to approximate both worlds, proving awareness about people actual behavior.
- It is the collection and analysis of machine sensed environmental data pertaining to human social behavior.
- It typically analyzes sensor data from mobiles, video cameras, satellites, etc

- Predictive patterns such as „honest signals" provide major factors in human decision making.
- Reality mining enables „big picture" of specific social contexts by aggregating and averaging the collected data
- It allows data/events correlation and consequently future occurrences extrapolation.

c. Context-Awareness

By assessing and analyzing visions and predictions on computing, devices, infrastructures and human interaction, it becomes clear that:

a. context is available, meaningful, and carries rich information in such environments,

b. that users" expectations and user experience is directly related to context,

c. acquiring, representing, providing, and using context becomes a crucial enabling technology for the vision of disappearing computers in everyday environments.

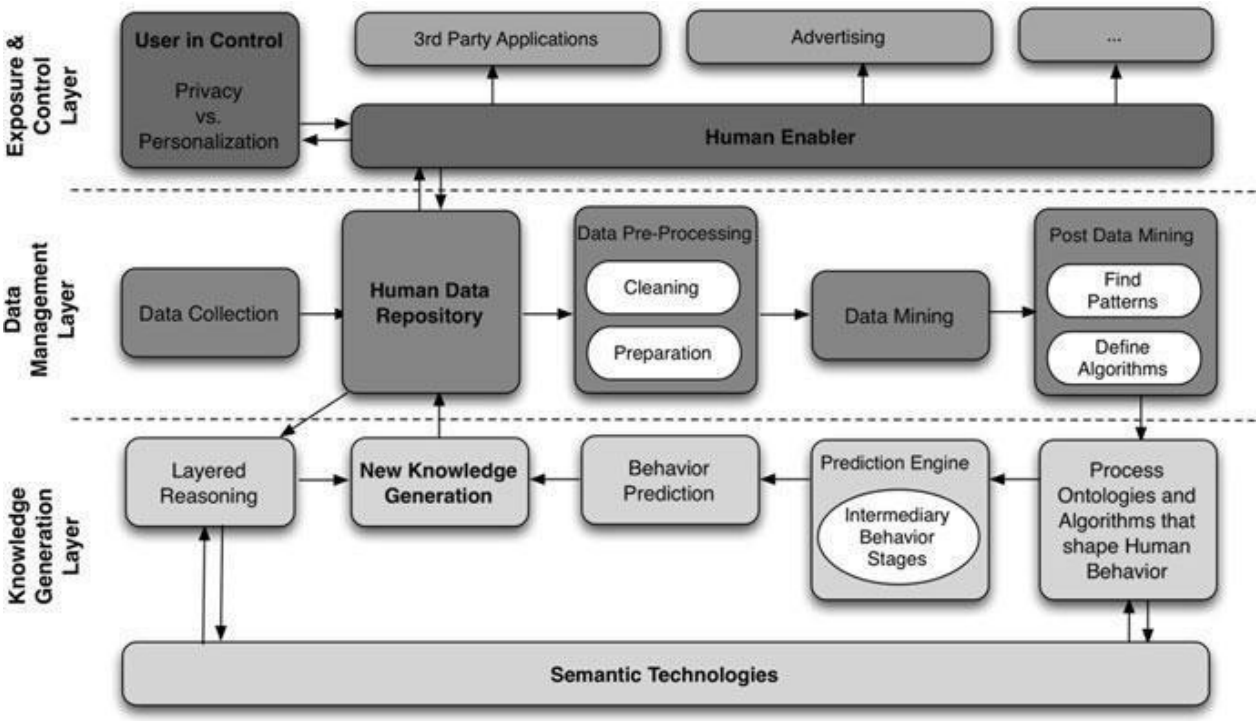
4.3.2 Architectural Framework and Methodology

To enable human behavior understanding and prediction, there are several independent but complementary steps that can be grouped into three different categories:

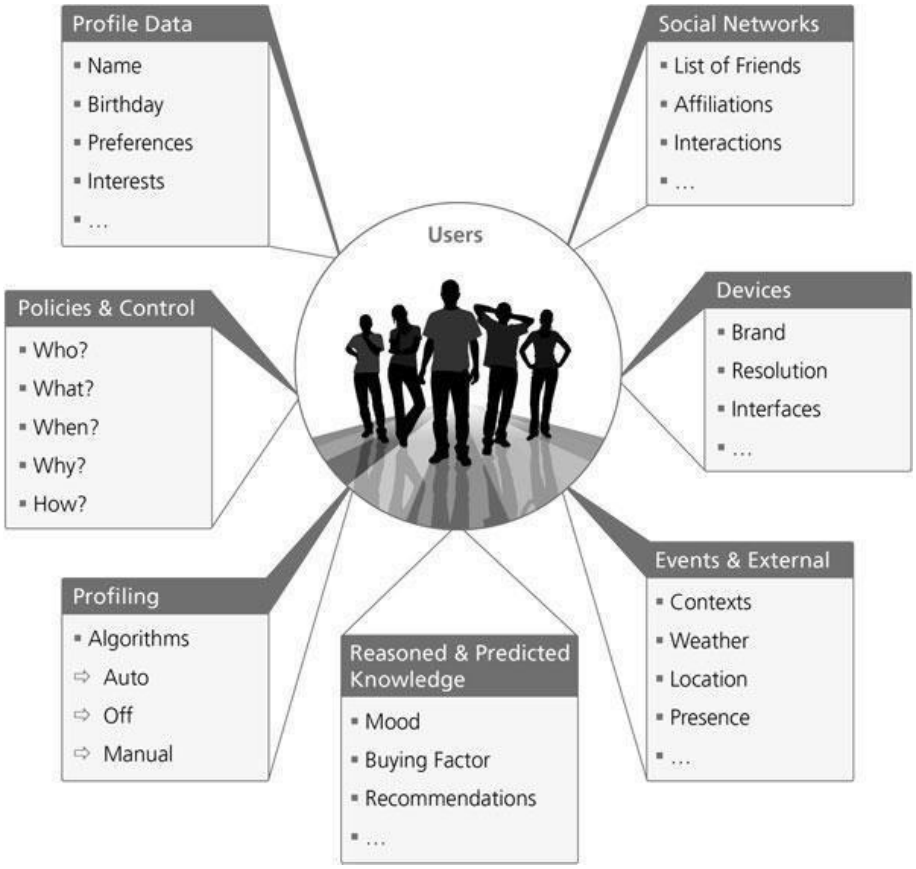
- ✓ Data Management
- ✓ New Knowledge Generation and
- ✓ Service Exposure and Control.

a. Data management

- This activity usually starts with data acquisition.
- Involves gathering of information from different information systems.
- Figure depicts these relationships as well as the sequence of activities involved.



- Figure below exemplifies the type of information that can be stored in the Human Data Repository



- Data is not usually captured without errors. Therefore it is necessary to preprocess it in advance before mining.
- Otherwise it would not be possible to correlate information correctly.
- Once this is done, data is mined by using two different approaches:
    - ✓ know statistical algorithms to help pattern recognition and consequent algorithmic modeling,
    - ✓ the opposite approach, where specific algorithms are designed to identify patterns in the data (this requires previous modeling).

Combining both, allows us to address the specifics of our applications, and at the same time, automatically detect new relevant correlations that might occur after a few iterations.

b. Knowledge Generation

- New information inference is based on user related data (called as context)
- Three different categories:
    - Real-time
    - Historical data
    - Reasoned context
- In Fig. below, there are several layers of abstraction in a context-aware system and any context-aware middleware or architecture must therefore be capable of building representations and models of these abstractions.
- However, these high-level abstractions can only be made from lower level context, which requires some form of context management function (performed by a Context Broker).
- In our case, this is performed at the **Human Data Repository**.
- The main context management features are context acquisition, context aggregation & fusion, context dissemination, discovery and lookup.
- In order to manipulate context information, it must be represented in some form that is compatible with the models that will be used in the reasoning and situation recognition processes.
- These models could be object oriented, ontological, rule based, logic based, based on semantic graphs or fuzzy logic sets.
- Reasoning mechanisms allow high-level context to be deduced or situations to be recognized that is output of one process can be used as an input to another.

- Reasoning is also used to check the consistency of context and context models.
- It is very important to stress that the prediction does not necessarily anticipates the user wishes or desires, but a possible future that could be interesting for the user.

c. Service Exposure and Control

- The third layer is divided into two main capabilities.
- The **first is user-centric** and relates to the ability of the user to stay in control of the whole scenario, enabling it to specify when, what, why, who, where and how the data isor can be accessed.
- Through the Human Enabler, users are able to influence the way their behavior is predicted, by controlling how there are being profiled (automatic, off, manually personalized).
- This is essential for establishing and managing trust and for safeguarding privacy, as well as for designing and implementing business security models and policies.
- The **second set of features is associated with the capacity of exposing this information** (both raw data and inferred one) to third party service providers (such as advertising agencies), through well defined web service interfaces.
- Besides exposing user related information, the human enabler allows data to be subscribed, syndicated or updated on request.

4.3.3. Innovations

The analysis of the first results indicated the following key findings:

- It is possible to infer user behavior based on user preferences, social networks and context-aware systems, with the help of reality/data mining techniques.
- Proximity and Similarity are great weight indicators for inferring influence and can be computed or calculated analytically.
- Both online and offline social networks have influence over a person"s behavior.
- User perceived QoE is improved as the methodology delivers personalization, contextualization, interactivity, adaptation and privacy.
- Users are willing to participate in their own profiling experience and the results are positive.

Applying these techniques into different fields of computer social sciences may have significant applicability in different parts of the value chain.

Examples:

- Infer and suggest missing information in users profile according to his/her peers contextual information.
- Understand how a specific user can be influenced by another user or community and vice versa.
- Understand how similar two users are, even if they do not have friends in common.

4.4 Privacy in Online Social Networks

- There is a dramatic growth in number and popularity of online social networks. There are many networks available with more than 100 million registered users such as Facebook, MySpace, QZone, Windows Live Spaces etc.
- People may connect, discover and share by using these online social networks. The exponential growth of online communities in the area of social networks attracts the attention of the researchers about the importance of managing trust in onlineenvironment.
- Users of the online social networks may share their experiences and opinions within the networks about an item which may be a product or service.
- **Collaborative filtering system** is the most popular method in recommender system.
    - The task is to predict the utility of items to a particular user based on a database of user rates from a sample or population of other users.
- Because of the different taste of different people, they rate differently according to their subjective taste.
- If two people rate a set of items similarly, they share similar tastes. In the recommender system, this information is used to recommend items that one participant likes, to other persons in the same cluster.
- Performs poor when there is insufficient previous common rating available betweenusers; known as cold start problem
- To overcome the cold start problem trust based approach to recommendation has emerged.

- This approach assumes a trust network among users and makes recommendations based on the ratings of the users that are directly or indirectly trusted by the target user.
- **Trust** could be used as supplementary or replacement of collaborative filtering system
- Trust and reputation systems can be used in order to assist users in predicting and selecting the best quality services
- Binomial Bayesian reputation systems normally take ratings expressed in a discretebinary form as either
    - positive (e.g. *good*) or
    - negative (e.g. *bad*).
- Multinomial Bayesian reputation systems allow the possibility of providing ratings with discrete graded levels such as e.g. *mediocre – bad –average – good – excellent*
- Trust models based on subjective logic are directly compatible with Bayesian reputation systems because a bi-jectivemapping exists between their respective trust and reputation representations.
- This provides a powerful basis for combining trust and reputation systems for assessing the quality of online services.
- Trust systems can be used to derive local and subjective measures of trust, meaning that different agents can derive different trust in the same entity.
- Reputation systems compute scores based on direct input from members in the community which is not based on transitivity
- Bayesian reputation systems are directly compatible with trust systems based on subjective logic, they can be seamlessly integrated. This provides a powerful and flexible basis for online trust and reputation management.

Online Social Networks

- A social network is a map of the relevant ties between the individuals, organizations, nations etc. being studied.
- With the evolution of digital age, Internet provides a greater scope of implementing social networks online. Online social networks have broader and easier coverage of members worldwide to share information and resources.
- The first online social networks were called UseNet Newsgroups. designed and built by Duke University graduate students Tom Truscott and Jim Ellis in 1979.

- Facebook is the largest and most popular online social network at this moment (www.insidefacebook.com).
- It had 350 million Monthly Active Users (MAU) at the beginning of January 2010. But it has been growing too fast around the world since then.
- As on 10 February 2010, roughly 23 million more people are using Facebook compared to 30 days ago, many in countries with big populations around the world. This is an interesting shift from much of Facebook"s international growth todate.
- Once Facebook began offering the service in multiple languages it started blowing up in many countries like Canada, Iceland, Norway, South Africa, Chile, etc.
- The United States is at the top with more than five million new users; it also continues to be the single largest country on Facebook, with 108 million MAU

Table a Top ten mostly visited social networks in Jan'09– based on MAU

Rank	Site	Monthly visit
1	facebook.com	1,191,373,339
2	myspace.com	810,153,536
3	twitter.com	54,218,731
4	flixster.com	53,389,974
5	linkedin.com	42,744,438
6	tagged.com	39,630,927
7	classmates.com	35,219,210
8	myyearbook.com	33,121,821
9	livejournal.com	25,221,354
10	imeem.com	22,993,608

4.5 Trust in Online Environment

- Trust has become important topic of research in many fields including sociology, psychology, philosophy, economics, business, law and IT.
- Trust is a complex word with multiple dimensions.
- Though dozens of proposed definitions are available in the literature, a complete formal unambiguous definition of trust is rare.
- Trust is used as a word or concept with no real definition.
- Trust is such a concept that crosses disciplines and also domains. The focus of definition differs on the basis of the goal and the scope of the projects.
- Two forms

- reliability trust or evaluation trust
- decision trust

- **Evaluation trust** can be interpreted as the reliability of something or somebody. It can be defined as the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends.
- The **decision trust** captures broader concept of trust. It can be defined as 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.

4.6 Trust Models Based on Subjective Logic

- Subjective logic is a type of probabilistic logic that explicitly takes uncertainty and belief ownership into account.
- 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.
- Subjective logic defines a trust metric called *opinion* denoted by $\omega_X^A = (\vec{b}, u, \vec{a})$ which expresses the relying party A‟s belief over a state space X.
  - Here $\vec{b}$ represents belief masses over the states of X, and *u* represent uncertainty mass where , u € $[\bar{0}, 1]$ and $\sum \bar{b} + u = 1$.
  - The vector $\vec{a} \in [0,1]$
  - represents the base rates over X, and is used for computing the probability expectation value of a state x.

- Binomial opinions are expressed as $\omega_x^A = (b, d, u, a)$ where d denotes disbelief in x. When the statement x for example says "*David is honest and reliable*", then the opinion can be interpreted as reliability trust in David.
- Let us assume that Alice needs to get her car serviced, and that she asks Bob to recommend a good car mechanic. When Bob recommends David, Alice would like to get

a second opinion, so she asks Claire for her opinion about David. This situation is illustrated in Fig. 4.6 a

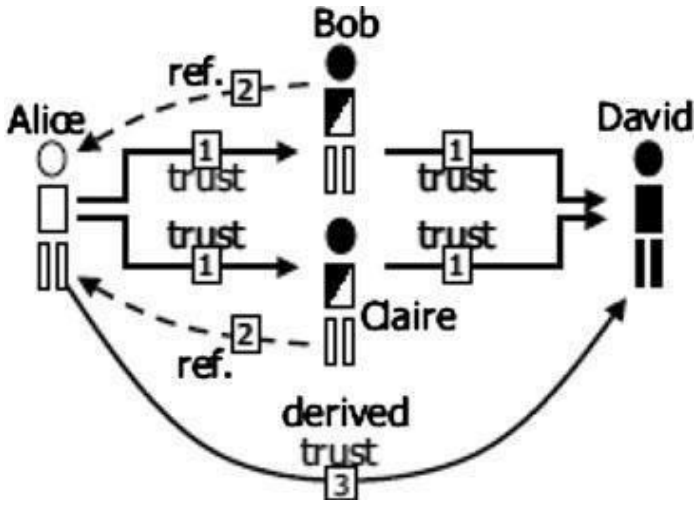


Fig. 4.6 a Deriving trust from parallel transitive chains

When trust and referrals are expressed as subjective opinions, each transitive trust path Alice →Bob→ David→ and Alice → Claire → David can be computed with the *transitivity operator*, where the idea is that the referrals from Bob and Claire are discounted as a function Alice"s trust in Bob and Claire respectively. Finally the two paths can be combined using the cumulative or averaging fusion operator. These operators form part of *Subjective Logic and semantic constraints* must be satisfied in order for the transitive trust derivation to be meaningful.

- This model is thus both belief-based and Bayesian.
- A trust relationship between A and B is denoted as [A:B]. The transitivity of two arcs is denoted as ":" and the fusion of two parallel paths is denoted as "◊". The trust network of Fig. 4.6 a can then be expressed as:

[A,D] = ([A,B]:[B,D]) ◊ ([A,C] : [C,D])

- The corresponding transitivity operator for opinions denoted as " " and the ⊗ corresponding fusion operator as "#. The mathematical expression for combining the opinions about the trust relationships of Fig. 4.6 a is then:

$$\omega_D^A = \left(\omega_B^A \otimes \omega_D^B\right) \oplus \left(\omega_C^A \otimes \omega_D^C\right)$$

- Arbitrarily complex trust networks can be analysed with TNA-SL which consists of a network exploration method combined with trust analysis based on subjective logic.
- The method is based on simplifying complex trust networks into a directed series parallel graph (DSPG) before applying subjective logic calculus.

4.7 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 network analysis using subjective logic (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.
    - In case of no explicit trust paths between two parties exist; subjective logic allows a level of trust to be derived through the default vacuous opinions.
    - TNA-SL is suitable for many types of trust networks.
    - Limitation : complex trust networks must be simplified to *series-parallel* networks in order for TNA-SL to produce consistent results.
    - The simplification consisted of gradually removing the least certain trust paths until the whole network can be represented in a series-parallel form.
    - As this process removes information it is intuitively sub-optimal.

4.7.1 Operators for Deriving Trust

- Subjective logic is a belief calculus specifically developed for modeling trust relationships.
    - 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*
- Expressed in the form of an ordered tuple

  $W_X^A = (b, d, u, a)$

  where b, d, and *u* represent belief, disbelief and uncertainty respectively where b, d, *u* ∈ [0, 1] and b+d+u = 1
- The base rate parameter a ∈ [0,1] represents the base rate probability in the absence of evidence, and is used for computing an opinion"s probability expectation value

$$E(W_X^A) = b + au$$

- A subjective opinion is interpreted as an agent A"s belief in the truth of statement x
- A"s opinion about x is denoted as W_X^A
- Subjective logic defines a rich set of operators for combining subjective opinions in various ways. Some operators represent generalizations of binary logic and probability calculus, whereas others are unique to belief calculus because they depend on belief ownership.

Transitivity is used to compute trust along a chain of trust edges. Assume two agents A and B where A has referral trust in B, denoted by W_B^A, for the purpose of judging the functional or referral trustworthiness of C.

In addition B has functional or referral trust in C, denoted by W_C^B. Agent A can then derive her trust in C by discounting B"s trust in C with A"s trust in B, denoted by $W_C^{A:B}$.

By using the symbol "⊗ to designate this operator, we define

$$\omega_C^{A:B} = \omega_B^A \otimes \omega_C^B \begin{cases} b_C^{A:B} = b_B^A b_C^B \\ d_C^{A:B} = b_B^A d_C^B \\ u_C^{A:B} = d_B^A + u_B^A + b_B^A u_C^B \\ a_C^{A:B} = a_C^B. \end{cases}$$

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

Let W_C^A and W_C^B be A"s and B"s trust in C respectively. The opinion $W_C^{A\diamond B}$ is then called the fused trust between W_C^A and W_C^B, By using the symbol "⊕ to designate this operator, we define

$$\omega_C^{A\diamond B} = \omega_B^A \oplus \omega_C^B \begin{cases} b_C^{A\diamond B} = \left(b_C^A u_C^B + b_C^B u_C^A\right) / \left(u_C^A + u_C^B - u_C^A u_C^B\right) \\ d_C^{A\diamond B} = \left(d_C^A u_C^B + d_C^B u_C^A\right) / \left(u_C^A + u_C^B - u_C^A u_C^B\right) \\ u_C^{A\diamond B} = \left(u_C^A u_C^B\right) / \left(u_C^A + u_C^B - u_C^A u_C^B\right) \\ a_C^{A\diamond B} = a_C^A. \end{cases}$$

The effect of the cumulative fusion operator is to amplify belief and disbelief and reduce uncertainty

4.7.2 Trust Path Dependency and Network Simplification

- Transitive trust networks can involve many principals
- Capital letters A, B, C and D will be used to denote principals.
- A single trust relationship can be expressed as a directed edge between two nodes that represent the trust source and the trust target of that edge.
- For example the edge [A, B] means that A trusts B. The symbol ":" is used to denote the transitive connection of two consecutive trust edges to form a transitive trust path.
- The trust relationships between four principals A, B, C and D connected serially can be expressed as:

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

- We will use the symbol "◊" to denote the graph connector. The "◊" symbol visually resembles a simple graph of two parallel paths between a pair of agents.
- In short notation, A"s combination of the two parallel trust paths from her to D is then expressed as:

([A,D]) = (([A,B] : [B,D]) ◊ ([A,C] : [C,D]))

- Trust networks can have dependent paths. This is illustrated on the left-hand side of Fig. 4.7.a. The expression for the graph on the left-hand side of Fig. 4.7.a would be:([A,D])

= (([A,B] : [B,D]) ◊ ([A,C] : [C,D]) ◊ ([A,B] : [B,C] : [C,D]))

- Trust network analysis with subjective logic may produce inconsistent results when applied directly to non-canonical expressions.

Fig. 4.7.a Network simplification by removing weakest path

It is therefore desirable to express graphs in a form where an arc only appears once. A canonical expression can be defined as an expression of a trust graph in structured notation where every edge only appears once.

Assuming that the path ([A,B]:[B,C]:[C,D]) is the weakest path in the graph on the left-hand side of Fig. 4.7.a, network simplification of the dependent graph would be to remove the edge [B,C] from the graph, as illustrated on the right-hand side of Fig. 4.7.a.

4.8 Trust Transitivity Analysis

Assume two agents A and B where A trusts B, and B believes that proposition x is true. Then by transitivity, agent A will also believe that proposition x is true. This assumes that B recommends x to A. In our approach, trust and belief are formally expressed as opinions. The transitive linking of these two opinions consists of discounting B"s opinion about x by A"s opinion about B, in order to derive A"s opinion about x. This principle is illustrated in Fig. 4.8.a below

Fig. 4.8.a Principle of trust transitivity

solid arrows - initial direct trust

dotted arrow - derived indirect trust

4.8.1 Uncertainty Favoring Trust Transitivity

A""s disbelief in the recommending agent B means that A thinks that B ignores the truth value of x. As a result A also ignores the truth value of x.

Uncertainty Favoring Discounting

Let A and B be two agents

where A"s opinion about B"s recommendations is expressed as

$$w_B^A = \{ b_B^A, d_B^A, u_B^A, a_B^A \}$$

let x be a proposition where B"s opinion about x is recommended to A with the opinion

$$w_x^B = \{ b_x^B, d_x^B, u_x^B, a_x^B \}$$

Let

$$w_x^{A:B} = \{ b_x^{A:B}, d_x^{A:B}, u_x^{A:B}, a_x^{A:B} \}$$

$$\begin{cases} b_x^{A:B} = b_B^A b_x^B \\ d_x^{A:B} = d_B^A d_x^B \\ u_x^{A:B} = d_B^A + u_B^A + b_B^A u_x^B \\ a_x^{A:B} = a_x^B \end{cases}$$

$w_x^{A:B}$ - the uncertainty favoring discounted opinion of A.

By using the symbol $\otimes$

$w_x^{A:B} = w_B^A \otimes w_x^B$

This operator is associative but not commutative. This means that the combination of opinions can start in either end of the path, and that the order in which opinions are combined issignificant. Figure below 4.8.b illustrates an example of applying the discounting operator for independent opinions, where

$w_B^A = \{0.1, 0.8, 0.1\}$ discounts $w_x^B = \{0.8, 0.1, 0.1\}$ to produce $w_x^{A:B} = \{0.08, 0.01, 0.91\}$

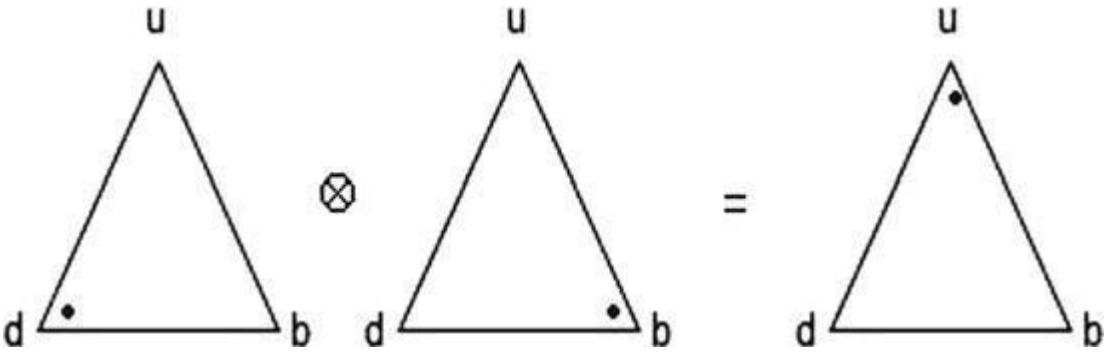


Fig. 4.8.b Example of applying the discounting operator for independent opinions

4.8.2 Opposite Belief Favoring

A‟s disbelief in the recommending agent B. A not only disbelieves in x to the degree that B recommends belief, but she also believes in x to the degree that B recommends disbelief in x, because the combination of two disbeliefs results in belief in this case.

Opposite Belief Favoring Discounting

Let A and B be two agents where A"s opinion about B"s recommendations is expressed as

$w_B^A = \{ b_B^A, d_B^A, u_B^A, a_B^A \}$

let x be a proposition where B's opinion about x is recommended to A with the opinion

$w_x^B = \{ b_x^B, d_x^B, u_x^B, a_x^B \}$

Let

$w_x^{A:B} = \{ b_x^{A:B}, d_x^{A:B}, u_x^{A:B}, a_x^{A:B} \}$

$$\begin{cases} b_x^{A:B} = b_B^A b_x^B + d_B^A d_x^B \\ d_x^{A:B} = b_B^A d_x^B + b_B^A d_x^B \\ u_x^{A:B} = u_B^A + (b_B^A + d_B^A) u_x^B \\ a_x^{A:B} = a_x^B \end{cases}$$

$w_x^{A:B}$ - opposite belief favoring discounted recommendation from B to A

By using the symbol $\otimes$

$w_x^{A:B} = w_B^A \otimes w_x^B$

This operator models the principle that "*your enemy's enemy is your friend*". It is doubtful whether it is meaningful to model more than two arcs in a transitive path with this principle. In other words, it is doubtful whether the enemy of your enemy's enemy necessarily is your enemy too.

4.8.3 Base Rate Sensitive Transitivity

Imagine a stranger coming to a town which is known for its citizens being honest. The stranger is looking for a car mechanic, and asks the first person he meets to direct him to a good car mechanic. The stranger receives the reply that there are two car mechanics in town, David and Eric, where David is cheap but does not always do quality work, and Eric might be a bit more expensive, but he always does a perfect job.

Translated into the formalism of subjective logic, the stranger has no other info about the person he asks than the base rate that the citizens in the town are honest. The stranger is thus ignorant, but the expectation value of a good advice is still very high.

Without taking a_B^A into account, the result of the definitions above would be that the stranger is completely ignorant about which of the mechanics is the best. An intuitive approach would then

be to let the expectation value of the stranger"s trust in the recommender be the discounting factor for the recommended (b_x^B, d_x^B) parameters.

Base Rate Sensitive Discounting

The base rate sensitive discounting of a belief

$w_x^B = \{ b_x^B, d_x^B, u_x^B, a_x^B \}$

by a belief

$w_B^A = \{ b_B^A, d_B^A, u_B^A, a_B^A \}$

Produces the transitive belief $w_x^{A:B} = \{ b_x^{A:B}, d_x^{A:B}, u_x^{A:B}, a_x^{A:B} \}$

$$\begin{cases} b_x^{A:B} = E\left(\omega_B^A\right) b_x^B \\ d_x^{A:B} = E\left(\omega_B^A\right) d_x^B \\ u_x^{A:B} = 1 - E\left(\omega_B^A\right)\left(b_x^B + d_x^B\right) \\ a_x^{A:B} = a_x^B \end{cases}$$

probability expectation value E

$w_B^A = \{ b_B^A + a_B^A u_B^A \}$

A safety principle could therefore be to only apply the base rate sensitive discounting to the last transitive link.

4.8.4 Mass Hysteria

Consider how mass hysteria can be caused by people not being aware of dependence between opinions. Let"s take for example; person A recommend an opinion about a particular statement x to a group of other persons. Without being aware of the fact that the opinion came from the same origin, these persons can recommend their opinions to each other as illustrated in Fig. 4.8.c

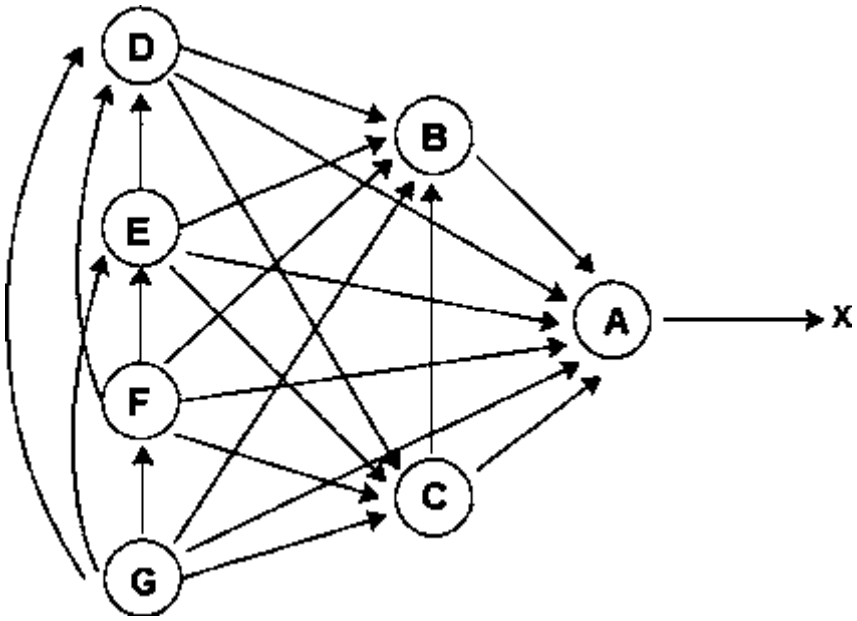


Fig. 4.8.c The effects of unknown dependence

The arrows represent trust so that for example B ->A can be interpreted as saying that B trusts A to recommend an opinion about statement x. The actual recommendation goes, of course, in the opposite direction to the arrows in Fig. 4.8.c.

If G assumes the recommended opinions to be independent and takes the consensus between them, his opinion can become abnormally strong and in fact even stronger than A"s opinion.

In order to reduce the size of the notation, the transitivity symbol ":" will simply be omitted, and the cumulative fusion symbol ◊ will simply be written as ",". Analyzing the whole graph of dependent paths, as if they were independent, will then produce:

$$\omega_x \left(\begin{array}{l} GA, GBA, GCA, GCBA, GDA, GDBA, GDCA, GDCBA, GEA, GEBA, GECA, \\ GECBA, GEDA, GEDBA, GEDCA, GEDCBA, GFA, GFBA, GFCA, GFCBA, \\ GFDA, GFDBA, GFDCA, GFDCBA, GFEA, GFEBA, GFECA, GFECBA, \\ GFEDA, GFEDBA, GFEDCA, GFEDCBA \end{array} \right) = (0.76, 0.11, 0.13, a)$$

When this process continues, an environment of self amplifying opinions, and thereby mass hysteria, is created.

4.9 Combining trust and reputation

4.9.1 The Dirichlet Reputation System

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.

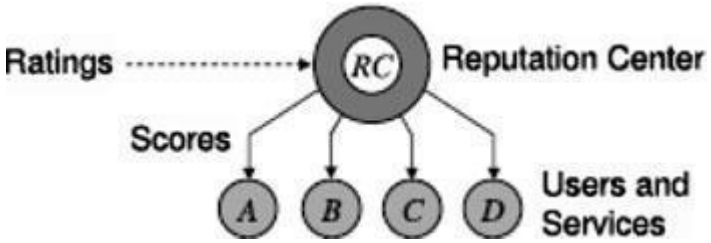


Figure 4.9.a Simple reputation system

Multinomial Bayesian systems are based on computing reputation scores by statistical updating of Dirichlet Probability Density Functions (PDF)

A posteriori (i.e. the updated) reputation score is computed by combining a priori (i.e. previous) reputation score with new ratings.

Agents are allowed to rate others agents or services with any level from a set of predefined rating levels

Reputation scores are not static but will gradually change

Let there be k different discrete rating levels. Let the rating level be indexed by i. The aggregate ratings for a particular agent can be expressed as

as: $\vec{R} = (\vec{R}(L_i)|i = 1 \ldots k).\psi\psi$

This vector can be computed recursively and can take factors such as longevity and community base rate into account.

$\vec{R}_y(L_i)\psi$ - aggregate rating of a particular level i for agent y

Before any ratings about a particular agent y have been received, its reputation is defined by common base rate

Ratings about particular agent are collected, the aggregate ratings can be computed recursively and derived scores will change accordingly.

The vector S is defined by

$$\vec{S}_y : \left(\vec{S}_y(L_i) = \frac{\vec{R}_y(L_i) + C\vec{a}(L_i)}{C + \sum_{i=1}^{k} \vec{R}_y(L_j)}; |i = 1 \ldots k\right).$$

The reputation score S is defined by

$$\sum_{i=1}^{k} \vec{S}(L_i) = 1\psi.$$

To express reputation score as a single value in some predefined interval. This can be done by assigning a point value to each rating level L :

$$\sigma = \sum_{i=1}^{k} v(L_i)\vec{S}(L_i).$$

A bijective mapping can be defined between multinomial reputation scores and opinions, which makes it possible to interpret these two mathematical representations as equivalent. The mapping can symbolically be expressed as:

$$\omega \leftrightarrow \vec{R}$$

Theorem: Equivalence Between Opinions and Reputations

Let w=(b,u,a) be an opinion, and R be a reputation, both over the same state space X so that the base rate „a" also applies to the reputation. Then the following equivalence

For $u \neq 0$*:*

$$\begin{cases} \vec{b}(x_i) = \frac{\vec{R}(x_i)}{C+\sum_{i=1}^{k}\vec{R}(x_i)} \\ u = \frac{C}{C+\sum_{i=1}^{k}\vec{R}(x_i)} \end{cases} \iff \begin{cases} \vec{R}(x_i) = \frac{C\vec{b}(x_i)}{u} \\ u + \sum_{i=1}^{k}\vec{b}(x_i) = 1 \end{cases}$$

For $u = 0$*:*

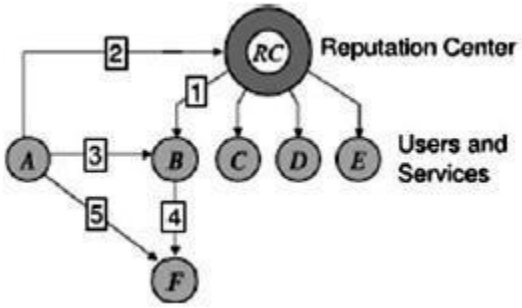
$$\begin{cases} \vec{b}(x_i) = \eta(x_i) \\ u = 0 \end{cases} \iff \begin{cases} \vec{R}(x_i) = \eta(x_i)\sum_{i=1}^{k}\vec{R}(x_i) = \eta(x_i)\infty \\ \sum_{i=1}^{k} m(x_i) = 1 \end{cases}$$

Multinomial aggregate ratings can be used to derive binomial trust in the form of an opinion. This is done by first converting the multinomial ratings to binomial ratings according to equationbelow and then apply to the above theorem.

The derived converted binomial rating parameters (r,s) are given by:

$$\begin{cases} r = \sigma\sum_{i=1}^{k}\vec{R}_y(x_i) \\ s = \sum_{i=1}^{k}\vec{R}_y(x_i) - r \end{cases}$$

Figure 4.9.b illustrates a scenario where agent A needs to derive a measure of trust in agent F .



Agent B has reputation score $\vec{R}_B^{RC}$ (arrow 1), and agent A has trust w_{RC}^A

RC in the Reputation Centre (arrow 2), so that A can derive a measure of trust in B (arrow 3). Agent B"s trust in F (arrow 4) can be recommended to A so that A can derive a measure of trust in F (arrow 5). Mathematically this can be expressed as:

$$\omega_F^A = \omega_{RC}^A \otimes \vec{R}_B^{RC} \otimes \omega_F^B$$

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.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF COMPUTING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT – V – SOCIAL NETWORK ANALYSIS – SITA3005

UNIT 5 GRAPH DATA IN THE REAL WORLD AND APPLICATION OF SOCIAL NETWORKS

Medium data - Tradition, Big Data, Small Data - Flat File Representations, Medium Data - Data Representation, Working with 2-Mode Data, Social Networks and Big Data, Big Data at work. Visualizing online social networks, Advances in Network Visualization - Elites, Communities and Influence, Applications of Social Network Analysis.

Medium data-Tradition

Relational SQL-based databases have traditionally been used to handle massive and dynamic datasets. Not only does this allow various components of the device to access the same data in one central location, but it also expands the size boundary to provide a lot of cheap hard drive space vs limited and costly RAM. Over the years, countless software programmes and applications have been created that link to a SQL database. There is almost always a SQL backend in whatever setting the analyst finds herself in. Of course, this includes Django, our personal favourite. It's only natural to handle social network data the same way we manage anything else: by saving and editing the graph directly in the database.

Big Data: The Future

Big Data is a collection of **data** that is **huge** in volume, yet growing exponentially with time. It is a **data** with so **large** size and complexity that none of traditional **data** management tools can store it or process it efficiently.

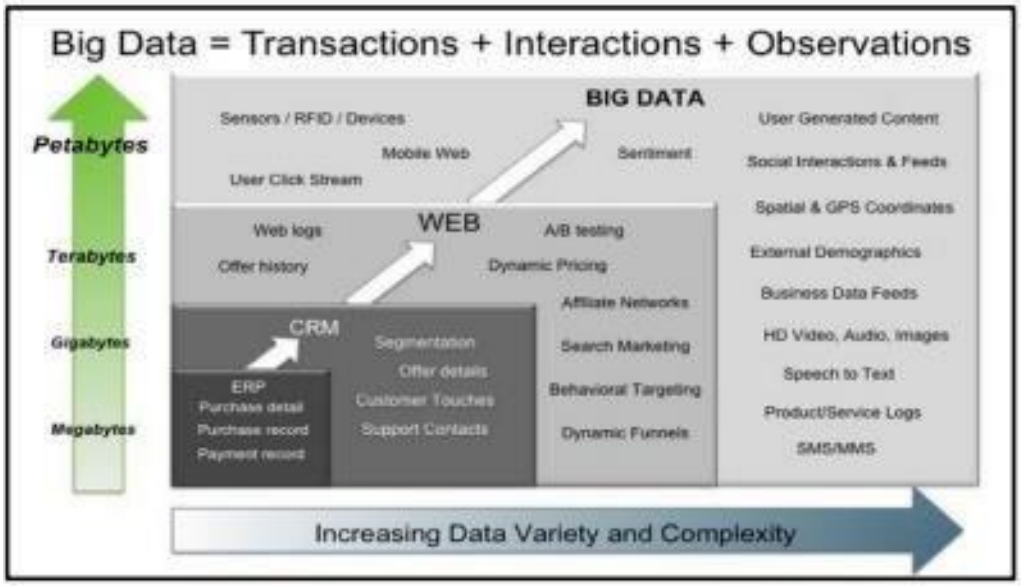


Figure 5.1: Big Data – Transactions, Interactions, Observations

Big Data Characteristics

The three Vs of Big data are Velocity, Volume and Variety.

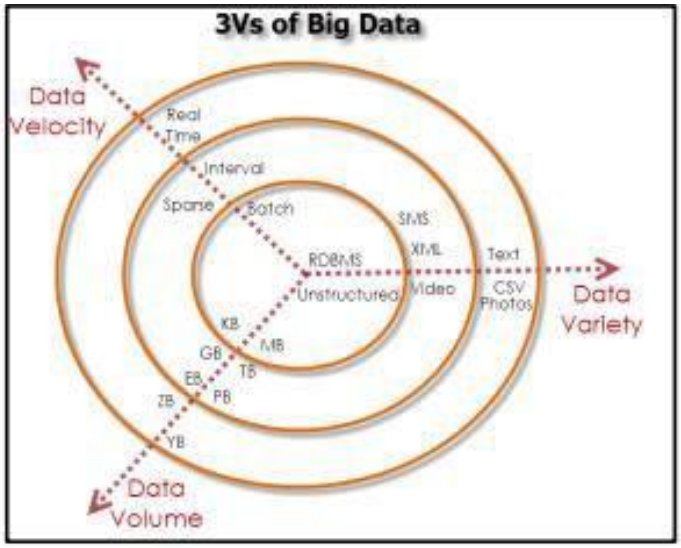


Figure 5.2: Big Data – Characteristics

VOLUME

The exponential growth in the data storage as the data is now more than text data. The data can be found in the format of videos, music's and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be re- evaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents Big data.

VELOCITY

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers

is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data.

VARIETY

Data can be stored in multiple format. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world have data in many different formats and that is the challenge we need to overcome with the Big Data.

Big Data Layout

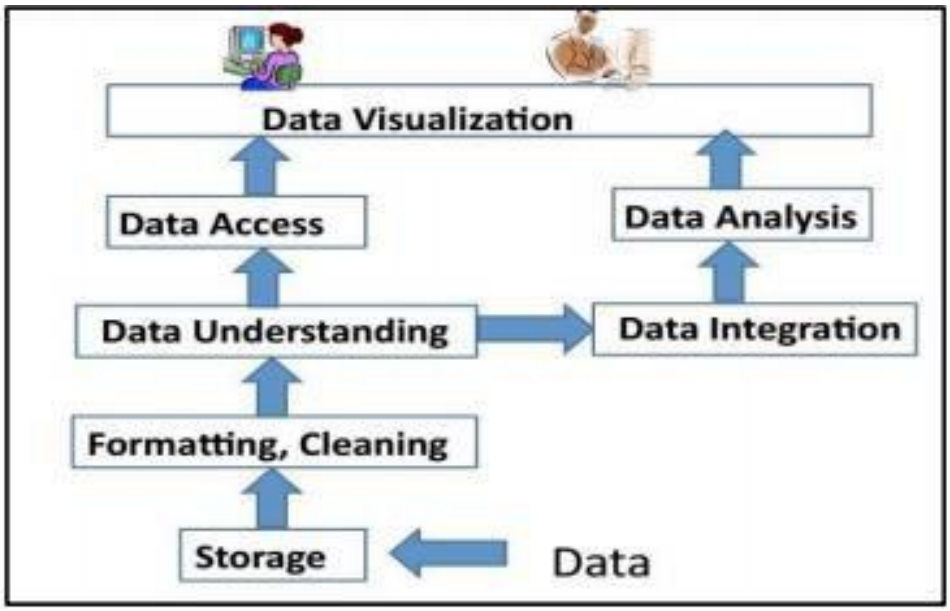


Figure 5.3: Big Data Layout

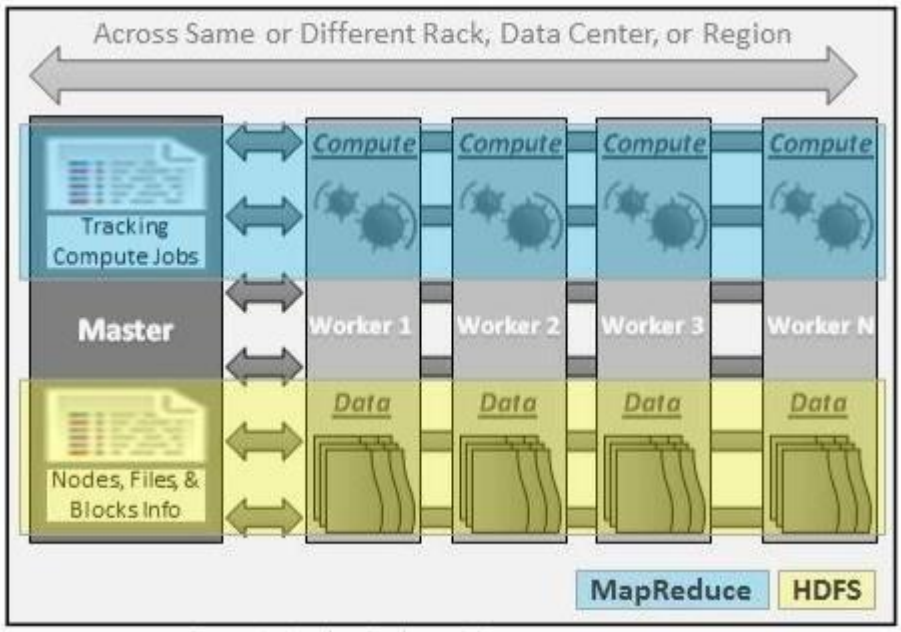


Figure 5.4: Big Data –Physical Architecture

1. APACHE HADOOP

Apache Hadoop is one of the main supportive element in Big Data technologies. It simplifies the processing of large amount of structured or unstructured data in a cheap manner. Hadoop is an open source project from apache that is continuously improving over the years. "Hadoop is basically a set of software libraries and frameworks to manage and process big amount of data from a single server to thousands of machines. It provides an efficient and powerful error detection mechanism based on application layer rather than relying upon hardware."

2. MAP REDUCE

MapReduce was introduced by google to create large amount of web search indexes. It is basically a framework to write applications that processes a large amount of structured or unstructured data over the web. MapReduce takes the query and breaks it into parts to run it on multiple nodes. By distributed query processing it makes it easy to maintain large amount of databy dividing the data into several different machines. Hadoop MapReduce is a software framework for easily writing applications to manage large amount of data sets with a highly faulttolerant manner. More tutorials and getting started guide can be found at Apache Documentation.

3. HDFS(Hadoop distributed file system)

HDFS is a java based file system that is used to store structured or unstructured data over large clusters of distributed servers. The data stored in HDFS has no restriction or rule to be applied, the data can be either fully unstructured of purely structured. In HDFS the work to make data senseful is done by developer's code only. Hadoop distributed file system provides a highly fault tolerant atmosphere with a deployment on low cost hardware machines. HDFS is now a part of Apache Hadoop project, more information and installation guide can be found at Apache HDFS documentation.

4. HIVE

Hive was originally developed by Facebook, now it is made open source for some time. Hive works something like a bridge in between sql and Hadoop, it is basically used to make Sql queries on Hadoop clusters. Apache Hive is basically a data warehouse that provides ad-hoc queries, data summarization and analysis of huge data sets stored in Hadoop compatible file systems. Hive provides a SQL like called HiveQL query based implementation of huge amountof data stored in Hadoop clusters.

5. PIG

Pig was introduced by yahoo and later on it was made fully open source. It also provides a bridge to query data over Hadoop clusters but unlike hive, it implements a script implementation to make Hadoop data access able by developers and business persons. Apache pig provides a high level programming platform for developers to process and analyses Big Data using user defined functions and programming efforts

Types of Data

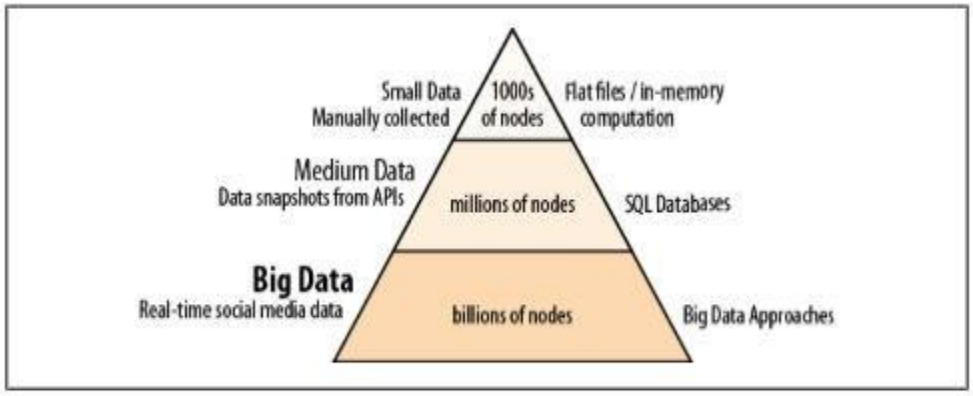


Figure 5.5 The realms of data sizes

Small data-flat files

- Flat files are great for quick analysis: they are portable and mostly human-readable.
- NetworkX is a tool to process graphs
- The graph we want to study has to fit entirely in memory.
- In general, memory usage is bounded by O(n)=n +n2=n(n+1).

NetworkX --supporting a number of different formats;

EdgeList Files

Perhaps the simplest way to store graph data is the Edgelist format. Its main advantage —besides simplicity—is the ability to easily import and export files into Excel or other spreadsheets.

- Edgelist files do not carry attribute data about nodes—but can carry arbitrary amounts of data concerning edges. The file format is exceedingly simple:
- <from_id> <to_id> <data1> <data2> ... <dataN>
- where from_id and to_id are string name or ID of the nodes that specify a graph edge, and columns <data1>...<dataN> are an arbitrary list of values that are assigned to this edge.
- freeform data can be kept in a Python dict format: <from_id> <to_id> <dict> foo bar {'weight':1,'color':'green' }

.net Format

_.net files represent a very expressing network data using ASCII text. This format was first used by a tool called Pajek, and is now something like a lingua franca of network data interchange. The files look something like this:

*Vertices 3

1 "Node1" 0.0 0.0 0.0 ic Green bc Brown

2 "Node2" 0.0 0.0 0.0 ic Green bc Brown

3 "Node3" 0.0 0.0 0.0 ic Green bc Brown

*Arcs 1 2 3 c Green 2 3 5 c Black *Edges 1 3 4 c Green The file is separated by sections, each section name starting with an asterisk (*):

*Vertices

- This section contains nodes' names and attributes.
- The header *Vertices is immediately followed by the number of vertices expected. Pajek expects this number to be accurate, although it is entirely optional elsewhere. Each of the lines contains the following columns:
- Numeric ID of the node (in sequence)
- Name of the node in quotes
- Node coordinates (x, y, z) and colors (background and foreground). These are non-zero only if the Pajek layout has been precomputed; they are optional for use in NetworkX .

*Arcs and *Edges

Arcs are directed edges, while Edges are undirected. If the graph is undirected, the Arcs section will not be present; if the graph is directed, the Edges section can be omitted. Each of the rows contains:

• from_id (a numeric ID of a node from the Vertices section)

• to_id (another numeric ID of a node)

• weight (value of the edge or arc)

• color (only if doing the layout in Pajek)

GML, GraphML, and other XML Formats

GML (Graph Modeling Language) and GraphML are two distant-cousin XML-based formats. A variety of tools support one or both of them—ranging from software for analysis of protein interactions, to project planning and supply chain planning systems. GraphML is more expressive, as it allows for hierarchical graphs, multigraphs, and other special cases .

A GraphML file looks something like this:

- <?xml version="1.0" encoding="UTF-8"?>

- <graphml> <key id="d0" for="node" attr.name="color" attr.type="string">yellow</key> <key id="d1" for="edge" attr.name="weight" attr.type="double"/> <graph id="G" edgedefault="undirected"> <node id="n0">

- <data key="d0">green</data> </node> <node id="n1"/> <node id="n2"> <data key="d0">blue</data> </node> <node id="n3"> <data key="d0">red</data> </node>

Ancient Binary Format—##h

- Files UCINET, a commercial Windows-based tool for SNA, has been considered “the Excel of SNA” for almost 20 years.

- It is a GUI-driven system encompassing hundreds of different modules, and is arguably the most comprehensive SNA tool on the planet.

“Medium Data”: Database Representation

A SQL-backed representation. Not only that, we will be able to add, remove, and modify nodes and edges in the stored graph directly in the database, without having to load the graphto memory

- def _prepare_tables(conn, name):

- c = conn.cursor()

- c.execute('drop table if exists "%s_edges"' % name)

- c.execute('drop table if exists "%s_nodes"' % name)

- c.execute('create table "%s_nodes" (node, attributes)' % name)

- c.execute('create table "%s_edges" (efrom, eto, attributes)' % name)

- conn.commit()

- c.close()

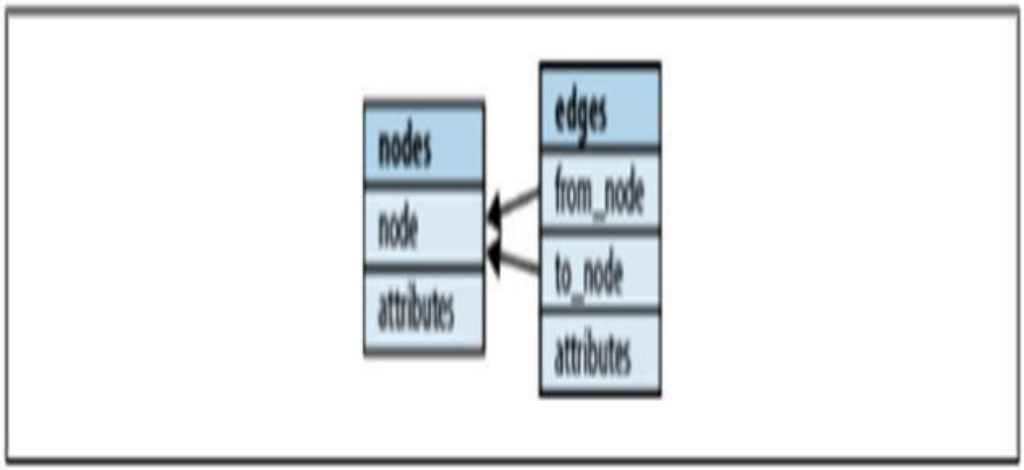


Figure 5.6 shows the database schema

Cursors

- Cursors are special database objects used to traverse tables.
- When a cursor executes a query, it returns another special object called a result set: a pointer into a set of results not yet returned by the server. The client then iterates over the result set, asking for one or several rows at a time, and the server only has to fetch them in small batches. This is partial lazy evaluation.

Nodes as Data, Attributes as ?

- In NetworkX a node is an object—it can be anything. Similarly, node/edge attributes are dictionaries of objects

Transactions

- Transactions are a safety feature that help prevent query errors from leaving data in a corrupt state, by executing a batch of operations atomically. Once a transaction is started, no changes are made to the actual data. It can then be either committed if no errors have occurred; or otherwise, rolled back. Committing a transaction commits the changes to the database, while a rollback discards them.

Names

- Aside from nodes and edges, NetworkX graphs have one more important piece of data: a name. In a database we can store multiple objects, so let's assume that a name uniquely identifies a graph.
- For simplicity, we'll use the graph name as part of the table name.

Functions and Decorators

- Before we proceed to methods that delete nodes and add/delete edges, we can observe an emerging pattern:
- 1. Get or create cursor
- 2. Execute a query
- 3. connection.commit()
- 4. cursor.close()

Function example

- def generalize(func):
- def generic(foo):
- cursor = connection.cursor()
- func(cursor, foo) # <- calling the wrapped function!
- cursor.close()

Decorator notation

- The technique of "wrapping" functions is commonly known as the Decorator Pattern. Python has special syntax for this:
- def add_node(cursor, node): cursor.execute('insert into my_nodes (node) value(?)', (foo,))
- add_node = generalize(add_node)

Working with 2 mode data

In social network analysis, 2-mode data refers to data recording ties between two sets of entities. In this context, the term "mode" refers to a class of entities – typically called actors, nodes or vertices – whose members have social ties with other members (in the 1- mode case) or with members of another class (in the 2-mode case). Most social network analysis is concerned with the 1-mode case, as in the analysis of friendship ties among a set of school children or advice- giving relations within an organization. The 2-mode case arises when researchers collect relations between classes of actors, such as persons and organizations, or persons and events. Forexample, a researcher might collect data on which students in a university belong to which

campus organizations, or which employees in an organization participate in which electronic discussion forums. These kinds of data are often referred to as affiliations. Co-memberships in organizations or participation in events are typically thought of as providing opportunities for social relationships among individuals (and also as the consequences of pre-existing relationships). At the same time, ties between organizations through their members are thoughtto be conduits through which organizations influence each other.

Perhaps the best known example of 2-mode network analysis is contained in the study of class and race by Davis, Gardner and Gardner (henceforth DGG) published in the 1941 book Deep South. They followed18 women over a nine-month period, and reported their participation in 14 events, such as a meeting of a social club, a church event, a party, and so on. Their original figure is shown in Figure 5.7

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×	.	×	×	.	.	.	.	.
2. Miss Laura Mandeville	×	×	×	.	×	×	×	×	.	.	.	.	.	.
3. Miss Theresa Anderson	.	×	×	×	×	×	×	×	×	.	.	.	.	.
4. Miss Brenda Rogers	×	.	×	×	×	×	×	×	.	.	.	.	.	.
5. Miss Charlotte McDowd	.	.	×	×	×	.	×	.	.	.	.	.	.	.
6. Miss Frances Anderson	.	.	×	.	×	×	.	×	.	.	.	.	.	.
7. Miss Eleanor Nye	.	.	.	.	×	×	×	×	.	.	.	.	.	.
8. Miss Pearl Oglethorpe	.	.	.	.	.	×	.	×	×	.	.	.	.	.
9. Miss Ruth DeSand	.	.	.	.	×	.	×	×	×	.	.	.	.	.
10. Miss Verne Sanderson	.	.	.	.	.	.	×	×	×	.	.	×	.	.
11. Miss Myra Liddell	.	.	.	.	.	.	.	×	×	×	.	×	.	.
12. Miss Katherine Rogers	.	.	.	.	.	.	.	×	×	×	.	×	×	×
13. Mrs. Sylvia Avondale	.	.	.	.	.	.	×	×	×	×	.	×	×	×
14. Mrs. Nora Fayette	.	.	.	.	.	×	×	.	×	×	×	×	×	×
15. Mrs. Helen Lloyd	.	.	.	.	.	.	×	×	.	×	×	×	.	.
16. Mrs. Dorothy Murchison	.	.	.	.	.	.	.	×	×	.	.	.	.	.
17. Mrs. Olivia Carleton	.	.	.	.	.	.	.	.	×	.	×	.	.	.
18. Mrs. Flora Price	.	.	.	.	.	.	.	.	×	.	×	.	.	.

Figure 5.7 shows the DGG women-by-events matrix.

DGG used the data to investigate the extent to which social relations tended to occur within social classes

A typical data matrix has two dimensions or ways, corresponding to the rows and columns of the matrix. The number of ways in a matrix X can be thought of as the number of subscripts needed to represent a particular datum, as in xij. If we stack together a number of similarly sized 2-dimensional matrices, we can think of the result as a 3- dimensional or 3-way matrix. The modes of a matrix correspond to the distinct sets of entities indexed by the ways. In the DGG dataset described above, the rows correspond to women and the columns to a different class of entities, namely events. Hence, the matrix has two modes in addition to two ways; it is 2-way, 2-mode. In contrast, a persons-by-persons matrix A, in which aij = 1 if person i is friends with person j, is a 2-way, 1-mode matrix, because both ways point to the same set of entities.

In a sense, what constitutes different modes is up to the researcher. If we collect romantic ties among a group of people of both genders, we could construct a 2-mode men-bywomen matrix X in which xij = 1 ifa romantic tie was observed between man i and woman j, and xij = 0 otherwise. Or, one could construct a larger 1-mode person-by-person matrix B also consisting of 1s and 0s in which it just happens that 1s only occur in cells where the row and column correspond to persons of different gender. Use of the men-bywomen matrix would imply that same-gender relations were impossible, whereas use of the person-by-person matrix would suggest that same-gender relations were logically possible, even if actually not observed.

Matrices recording relational information such as romantic ties can be represented as mathematical graphs as well. A graph G(V,E) consists of a set of nodes or vertices V together with a set of lines or edges E that connect them. An edge is simply an unordered pair of nodes (u,v). (In directed graphs or digraphs we use ordered pairs to indicate direction of the tie.) To indicate a tie between two nodes u and v, we simply include the pair (u,v) in the set E. The number of nodes in a graph is denoted by |V| or n.

A bipartite graph is a graph in which we can partition all nodes into two sets, V1 and V2, such that all edges include a member of V1 and a member of V2. The number of nodes in each vertex set is denoted n1 and n2, respectively.

Two-Mode Data in Social Network Analysis

Most social networks are conceived of as relations among a set of nodes, and therefore represented as a1-mode matrix (typically of 1s and 0s) or a simple graph or digraph. For example, we might collect data on who is friendly with whom within an organization, or who injects drugs with whom in a neighborhood. However, 2-mode data are common in social network contexts as well. Typical examples include, actor-by-event attendance (as in the DGG data), actor by group membership (such as managers sitting on corporate boards), and actor by trait possession (such as adjective checklist data), and actor by object possession (such as material style of life scales in which inventories are made of household possessions). In many cases when 2-mode data are collected, the analytical interest is focused on one mode or the other. For example, in the DGG dataset, person-by-event attendances were collected in order to understand social relations among the women, specifically, whether women tended to have social relations primarily within their own social classes.

In the interlocking directorate literature, membership of executives on corporate boards is collected mainly in order to understand how corporations are intertwined, and how the structure of this connectivity affects corporate control of society. However, it can also occur that neither mode dominates our analytical focus and the primary interest is in the correspondence of one mode to the other. For example, a university might ask its faculty which courses they prefer to teach. Here, the objective is typically not to understand how faculty are related to each other through courses, nor how courses are related via faculty, but in the optimal assignments of persons to courses so that courses are staffed and faculty are not complaining.

Social networks and big data

Our data transmission, processing, and storage capacities have become practically limitless and inexpensive. As they currently exist, social networks are essential to us in one primary way: as massive data collectors of human behaviour patterns, enthusiastically powered by the examined themselves, and functioning on a global scale with near-instant response times. The end result is massive amounts of data.

NOSQL

NoSQL is a class of databases that is a departure (sometimes radical), from the classic RDBMS. NoSQL databases typically do not support a query language and lack a fixed schema. They exist in different variations: document stores, key-value stores, object stores, graph databases, and so on. Each variation is designed differently, with a particular context in mind. They are important to us for three common reasons: structure, size, and computation—the same reasons why relational databases are often ill-suited for our purposes. Here are some NoSQL examples:

BigTable

Google's implementation of a document store, designed to scale across assorted inexpensive hardware. They use it to store the whole Internet. The whole Internet.

HBase

Open source clone of BigTable, by the Apache Software Foundation. A work in progress, but moving quickly.

Hadoop

Not a database itself, but the underlying framework for distributed storage and processing, by ASF. The Hadoop project is very impressive and is one of the established standards for Big Data and cloud computing.

MongoDB

A prominent scalable high-performance document store, with support for indexes, queries and more. Uses JSON documents with dynamic schemas.

CouchDB

Another distributed document-oriented database, by ASF. Simple to install and operate, and supports MapReduce style indexing (more on MapReduce below).

Neo4j

A high-performance graph database written in Java.

Hive

A SQL-like database that can operate on plan text files in a variety of formats.

Structural Realities

The world is a fickle and ever-changing place. The structural coherence of Big Data is not always assured, owing to the fact that the data comes from a variety of different sources. Social media platforms are no exception. And more so, as the market progresses, the data structure evolves at a breakneck pace. Social networking sites are constantly updating their functionality and data standards. We must contend with not only inconsistencies between providers, but also changes in data from the same provider over time. Because of this structure flux, any attempt to coerce, say, a Twitter stream into a relational schema is futile.

Plain Text

The plain-text aspect is of great practical value to us. Suppose we were collecting Twitter data in the format shown above, from an HTTP stream.

With a SQL database, a data ingestion flow may look like this:

1. First, learn the data structure and possibly design a fitting schema.

2. Read the stream. For every record: a. Parse the record. b. Map record fields to an INSERT query. c. Execute query, handle unique constraint violations. d. Optionally update other tables.

With a text-file-based NoSQL database, the flow becomes this:

1. Read stream.

2. Write record to file.

Computational Complexities

Distributed computing is a technique of breaking a big task into smaller subtasks, distributing them across multiple machines, and combining the results. One can have a hardware array in- house or turn to the cloud and rent virtual machines en masse, on demand, for pennies per hour (each). Horizontal scaling means NoSQL databases are perfectly suited for the cloud.

Big Data at Work

Distributed computing

Distributed computing is a model in which components of a software system are shared among multiple computers. Even though the components are spread out across multiple computers, they are run as one system. This is done in order to improve efficiency and performance.

Hadoop, S3 and Map Reduce

A distributed cluster of computational nodes needs some way to access the source data. Hadoop by ASF makes that easy with HDFS (Hadoop file system) and features like data streaming, for chaining processes. Hadoop automatically distributes chunks of source data between nodes, controls their operation, and monitors their progress via a supplied Web interface

Amazon S3 (Simple Storage Service) is an online storage service by Amazon. It is distributed and scalable, though the implementation is kept under wraps and to the user it appears as a gigantic disk in some contexts and as a key-value store in others (the key is the filename and the value is the file content). S3 objects can be accessed over HTTP, making it a great Web storage medium. More interestingly, S3 can be accessed as a Hadoop file system.

MapReduce is a computational framework invented by Google and supported by Hadoop. A MapReduce job consists, not surprisingly, of a map step and a reduce step— both names of common functions in functional programming, though not exactly the same. Map step takes the source input, one record at a time, and outputs some meaningful data about it. Multiple copies of the map step work each on a different subset of the source data. Reduce step takes the output of several (or all) map steps and combines them.

Hive

Hive is another project by ASF. It is unique for its effort to bring SQL back into a NoSQL world. Hive operates by transforming SQL queries into MapReduce jobs and presenting their results back in SQL form. Hive has the important ability to read and write flat files as SQL tables, using what they call a SerDe (serializer/deserializer) to map file contents to columns. There are SerDes provided for CSV, JSON, and regexp formats, and more can be added. The best part is that it can read a collection of files on S3.

2-Mode Networks in Hive- a recipe

```
create external table hashtags ( tweet_id string, text string, indices string )

row format serde 'com.amazon.elasticmapreduce.JsonSerde'

with serdeproperties ('paths'='id, entity.text, entity.indices')

location 's3://our-json-data/output/hashtags/';
```

2-mode networks are just as easy with Hive as any other SQL server.

Using this table as an example, this will produce an edge list, weighted by cooccurrence counts, of the hashtag network:

hive> select x.text, y.text, count(x.tweet_id) from hashtags x full outer join hashtags y on (x.tweet_id=y.tweet_id) where x.text != y.text group by x.text, y.text;

Visualising Online Social networks

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 at least the 1930s. Linton Freeman documents the history of social network visualization within sociological research, providing examples of the ways in which spatial position, color, size, and shape can all be used to encode information . For example, networks can be arranged on a map to represent the geographic distribution of a population. Alternatively, algorithmically generated layouts have useful spatial properties: a force-directed layout can be quite effective for spatially grouping connected communities, while a radial layout intuitively portrays network distances from a central actor. Color, size, and shape have been used to encode both topological and non-topological properties such as centrality, categorization, and gender.

Vizster –Visualisation tool

Vizster was to build a visualization system that end-users of social networking services could use to facilitate discovery and increased awareness of their online community. We wanted to support the exploratory and playful aspects of Friendster while also giving users easier access to search and group patterns. While users regularly explored the network on Friendster, the linear format limited such explorations. This led us to develop richer network views and exploratory tools, while maintaining a local orientation. We also learned that the use of imagery was indispensable for identifying people and establishing a presentation of self, and so must play a central role in the visualization. In addition to helping support the current practices, we wanted to make sure that Vizster did not eliminate the data that helped users get a sense of people through their profiles. One example is the use of re-appropriated profile fields (e.g., inverting ages to identify teenagers) for coded communication within a subpopulation. For this reason, we realized that we must make searchable profile data very present and accessible in the visualization. These goals position Vizster differently from

traditional social network visualizations used as analysis tools by social science researchers. The following description includes the implications this approach has had for our design decisions, both in terms of presentation and the level of technical sophistication exposed by the visualization. Vizster presents social networks using a familiar node-link representation, where

nodes represent members of the system and links represent the articulated "friendship" links between them (Figures 5.8). In this view, network members are presented using both their self-provided name and, if available, a representative photograph or image. The networks are presented as egocentric networks: networks consisting of an individual and their immediate friends. Users can expand the display by selecting nodes to make visible others' immediate friends as well. To the right of the network display is a panel presenting a person's profile. As discussed later, the profile panel also provides direct manipulation searches over profile text.

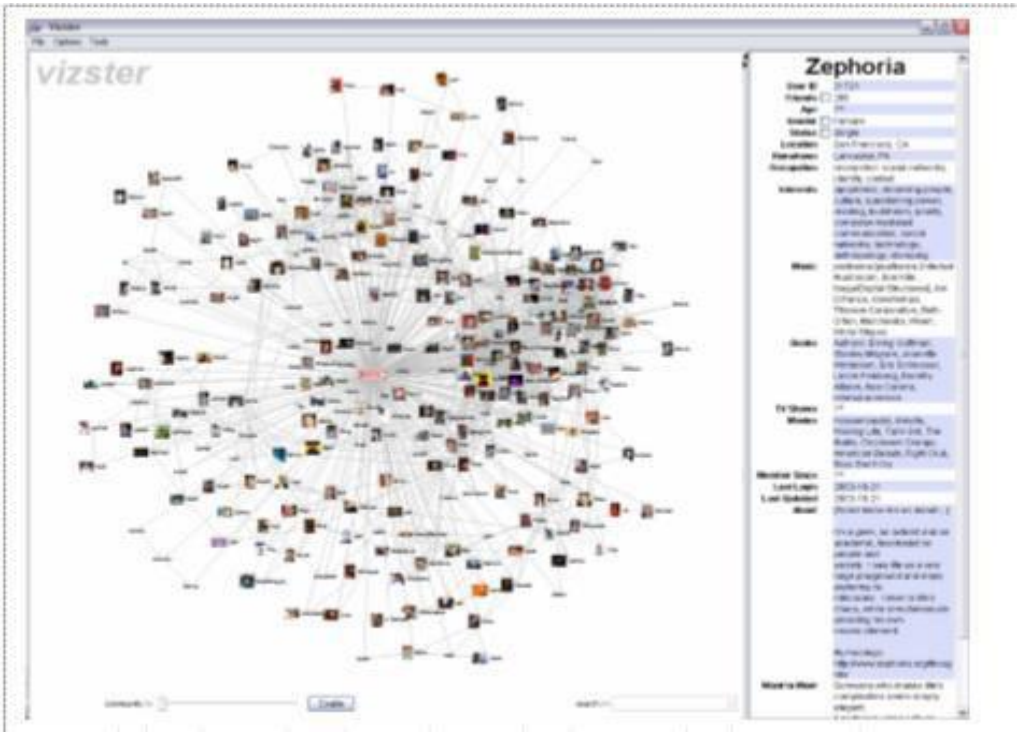


Figure 5.8 shows the screen short of Vizster visualization system

Applications of social network analysis

Social network analysis is used extensively in a wide range of applications and disciplines. Some common network analysis applications include data aggregation and mining, network propagation modeling, network modeling and sampling, user attribute and behavior analysis, community-maintained resource support, location-based interaction analysis, social sharing and filtering, recommender systems development, and link prediction and entity resolution.[56] In the private sector, businesses use social network analysis to support activities such as customer interaction and analysis, information system development analysis,[57] marketing, and business intelligence needs (see social media analytics). Some public sector uses include development of leader engagement strategies, analysis of individual and group engagement and media use, and community-based problem solving.

Security applications

Social network analysis is also used in intelligence, counter-intelligence and law enforcement activities. This technique allows the analysts to map covert organizations such as an espionage ring, an organized crime family or a street gang. The National Security Agency (NSA) uses its electronic surveillance programs to generate the data needed to perform this type of analysis on terrorist cells and other networks deemed relevant to national security. The NSA looks up to three nodes deep during this network analysis. After the initial mapping of the social network is complete, analysis is performed to determine the structure of the network and determine, for example, the leaders within the network This allows military or law enforcement assets to launch capture-or-kill decapitation attacks on the high-value targets in leadership positions to disrupt the functioning of the network. The NSA has been performing social network analysis on call detail records (CDRs), also known as metadata, since shortly after the September 11 attacks.

Textual analysis applications

Large textual corpora can be turned into networks and then analysed with the method of social network analysis. In these networks, the nodes are Social Actors, and the links are Actions. The extraction of these networks can be automated by using parsers. The resulting networks, which can contain thousands of nodes, are then analysed by using tools from network theory to identify the key actors, the key communities or parties, and general properties such as robustness or structural stability of the overall network, or centrality of certain nodes.This automates the approach introduced by Quantitative Narrative Analysis, whereby subject-verb-object triplets are identified with pairs of actors linked by an action, or pairs formed by actor-object. In other approaches, textual analysis is carried out considering the network of words co-occurring in a text (see for example the Semantic Brand Score). In these networks, nodes are words and links among them are weighted based on their frequency of co-occurrence (within a specific maximumrange).

Internet applications

Social network analysis has also been applied to understanding online behavior by individuals, organizations, and between websites.Hyperlink analysis can be used to analyze the connections between websites or webpages to examine how information flows as individuals navigate the web. The connections between organizations has been analyzed via hyperlink analysis to examine which organizations within an issue community.

Social media internet applications

Social network analysis has been applied to social media as a tool to understand behavior between individuals or organizations through their linkages on social media websites such as Twitter and Facebook.

In computer-supported collaborative learning

One of the most current methods of the application of SNA is to the study of computer-supported collaborative learning (CSCL). When applied to CSCL, SNA is used to help understand how learners collaborate in terms of amount, frequency, and length, as well as the quality, topic, and

strategies of communication. Additionally, SNA can focus on specific aspects of the network connection, or the entire network as a whole. It uses graphical representations, written representations, and data representations to help examine the connections within a CSCL network When applying SNA to a CSCL environment the interactions of the participants are treated as a social network. The focus of the analysis is on the "connections" made among the participants – how they interact and communicate – as opposed to how each participant behaved on his or her own.