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.

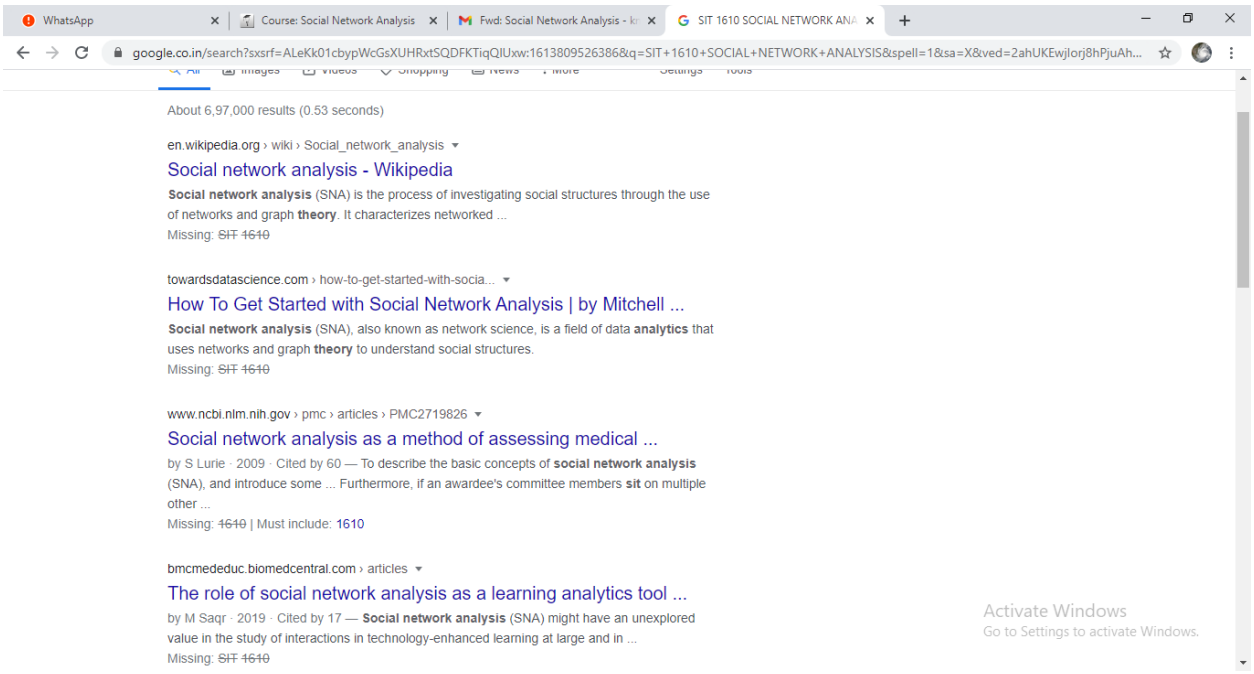
 This means adaptation to our primary interface to the vast information that constitutes the Web: the search engine.

 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 top ten 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.



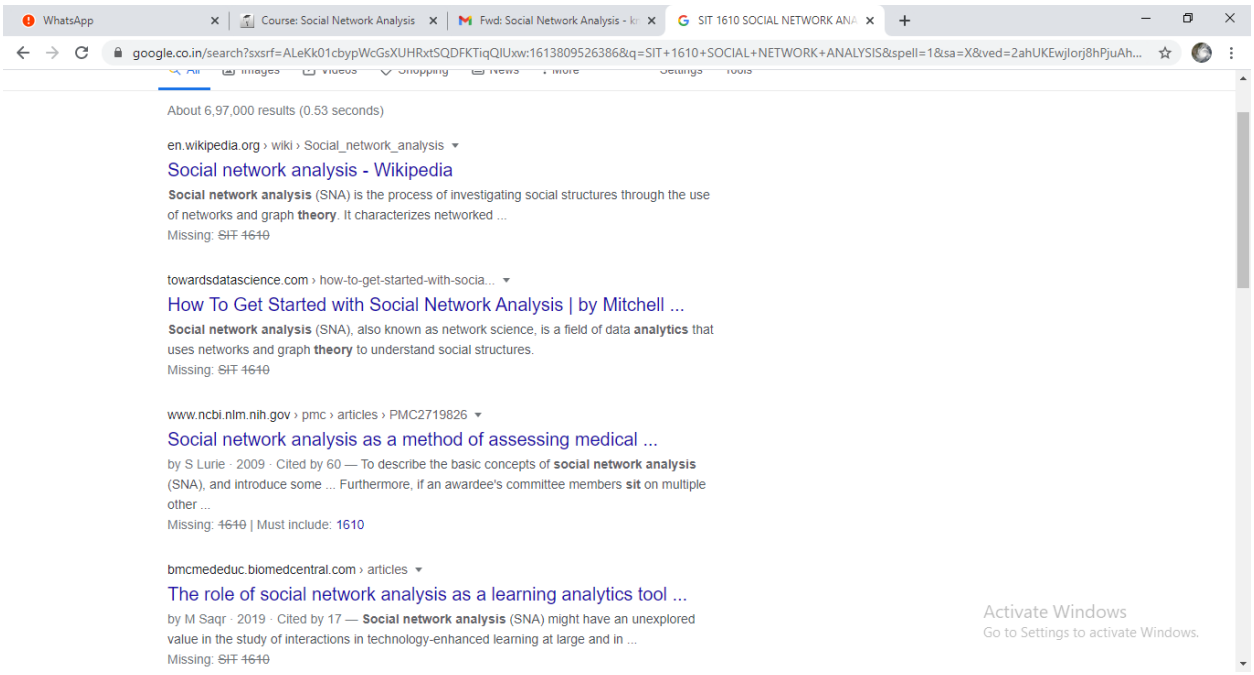


Fig.1 Search results for the keyword SIT1610 *Social network analysis* using Google

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 on the 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.

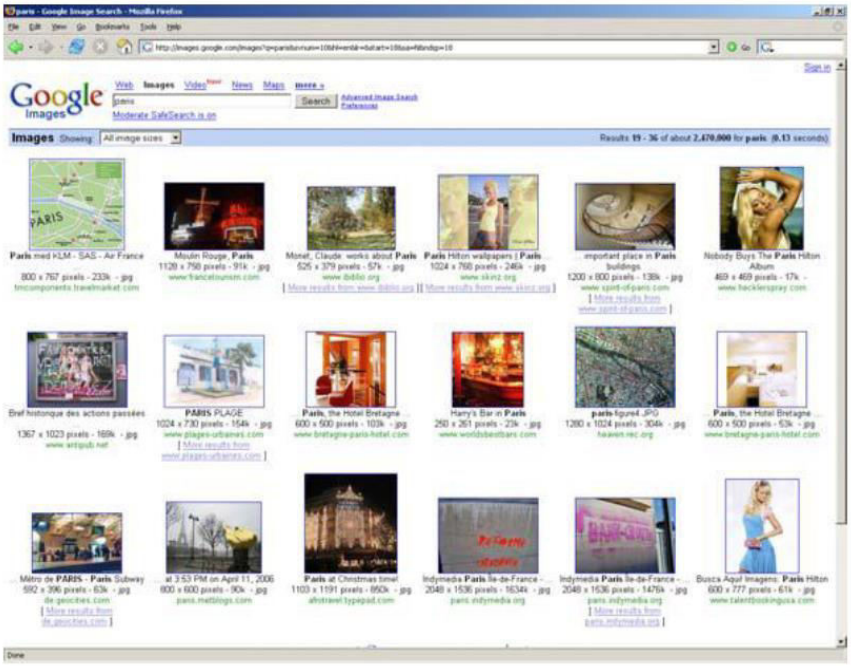
Even if an algorithm can determine that the page describes a music player, information about the product is very difficult to spot.

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 described in formal rules (**e.g. if A is true, B should follow**). But formal knowledge typically captures only the 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.

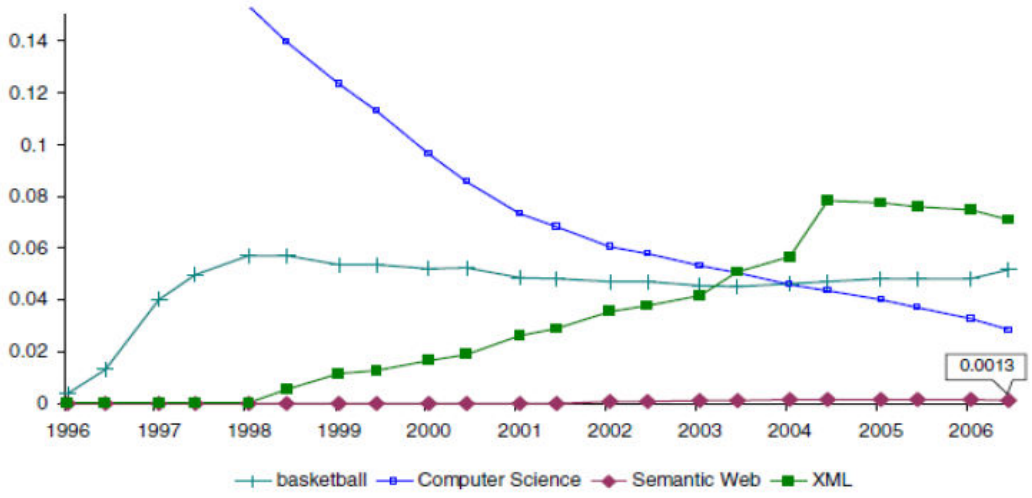


Fig2. Number of webpage with the terms *basketball*, *Computer Science*, and *XML* over time and as a fraction of the number of pages with the term *web*.

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.

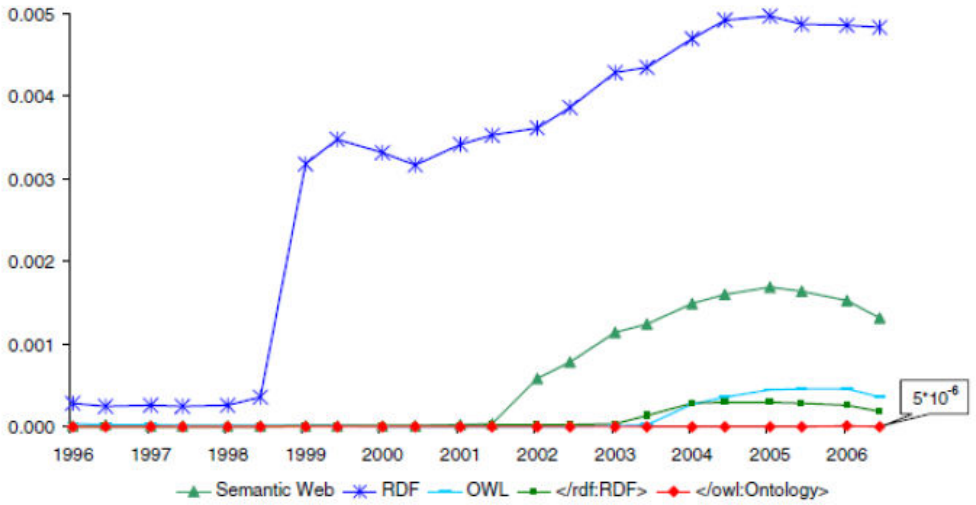


Fig3. Number of WebPages with the terms RDF, OWL and the number of ontologies in RDF or OWL over time. Again, the number is relative to the number of pages with the term web.

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.

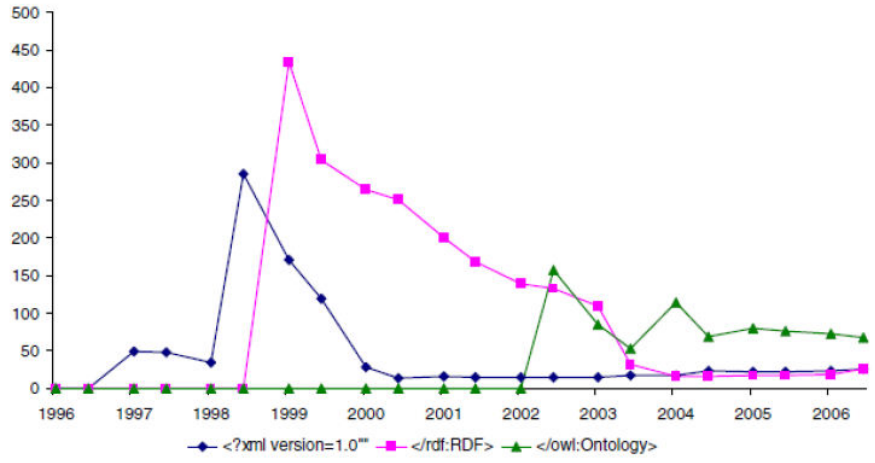


Fig.4 The hype cycle of Semantic Web related technologies as shown by the number of web pages about a given technology relative to its usage

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 and evolves 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

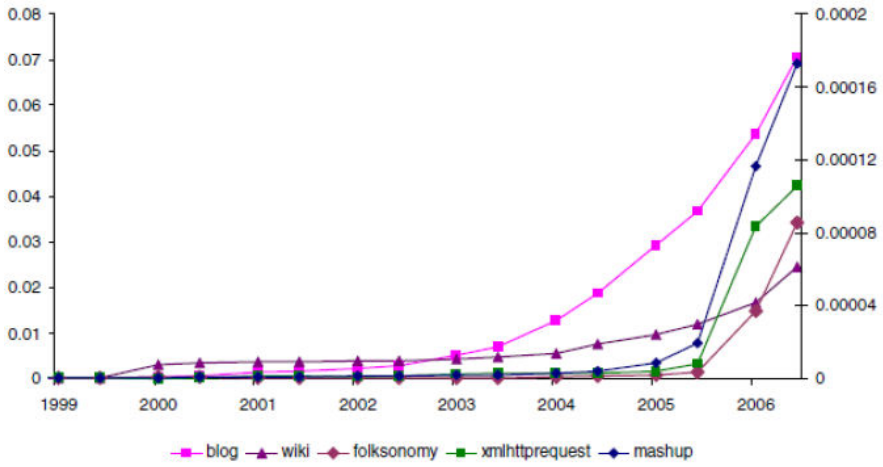


Fig.5 Development of the social web.

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 combining data 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 network analysis 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 part of 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 representations can 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 and contributes 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.

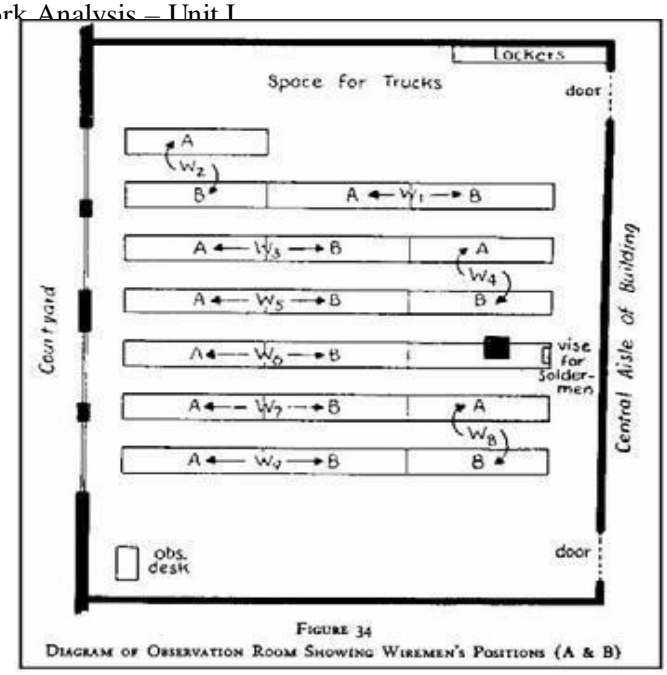


FIGURE 34
DIAGRAM OF OBSERVATION ROOM SHOWING WIREMEN'S POSITIONS (A & B)

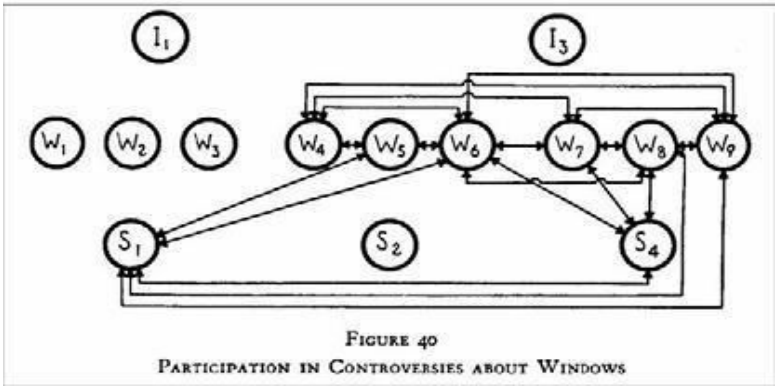


FIGURE 40
PARTICIPATION IN CONTROVERSIES ABOUT WINDOWS

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:

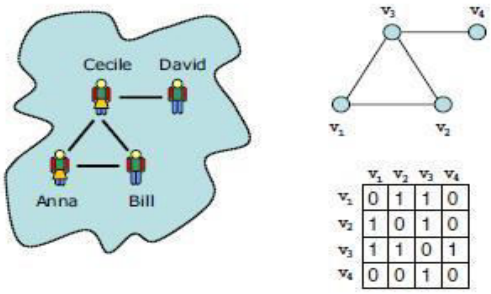


Fig. 7 Most network analysis methods work on an abstract, graph based representation of real world networks.

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 path that 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.

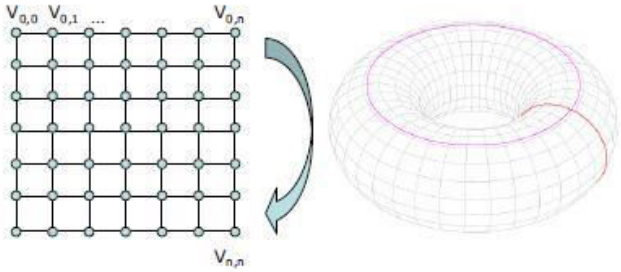
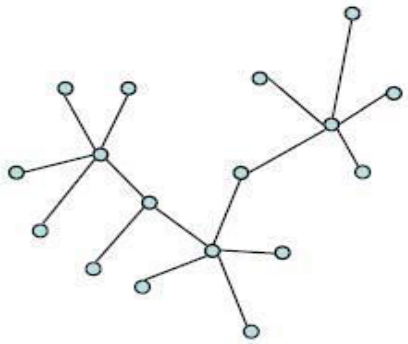


Fig.8 The 2D lattice model of networks (left). By connecting the nodes on the opposite borders of the lattice we get a toroidal lattice (right).

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.

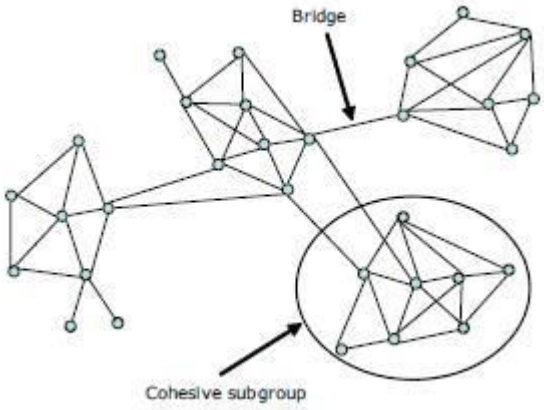


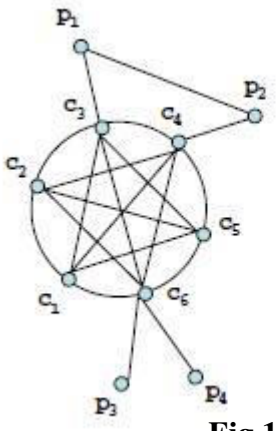
Fig.10 Most real world networks show a structure where densely connected subgroups are linked together by relatively few bridges

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.

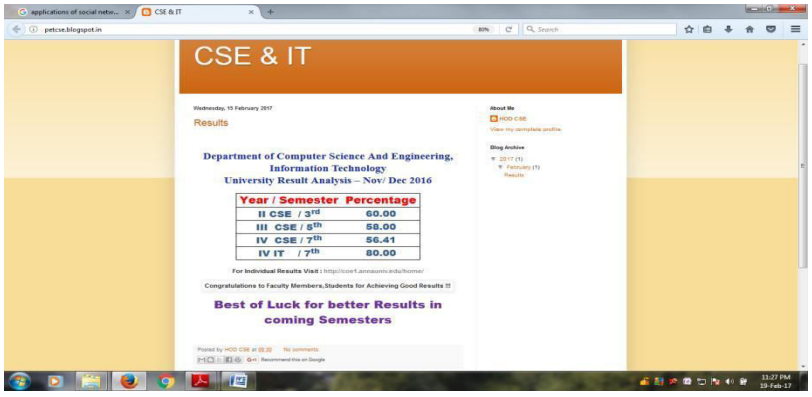
The studies of electronic communication networks based on email data are limited by privacy concerns. For example, in the HP case the content of messages had to be ignored by the researchers and the data set could not be shared with the community.

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.

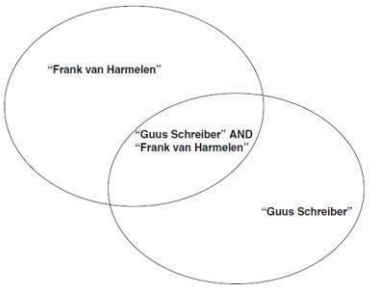
On the other hand, extracting relationships based on co-occurrence of the names of individuals or institutions requires web mining as names are typically embedded in the natural text of web pages.

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.

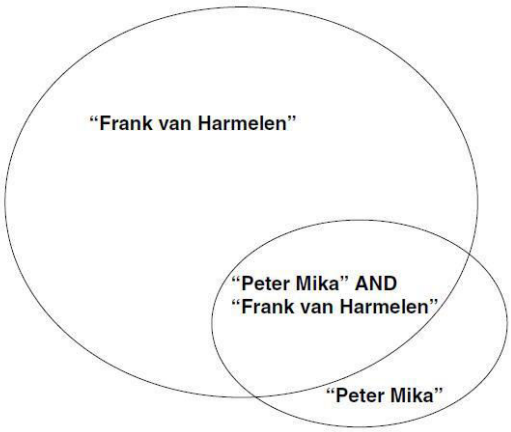
The resulting value of tie strength is a number between zero (no co-occurrences) and one (no separate mentioning, only co-occurrences). If this number has exceeded a certain fixed threshold it was taken as evidence for the existence of a tie.

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

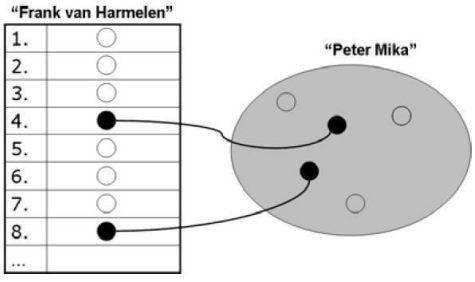
A Hybrid Approach Using PSO and K-Means for Semantic Clustering of Web Documents. J. Web Eng. 12(3&4): 249-264 (2013)

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.

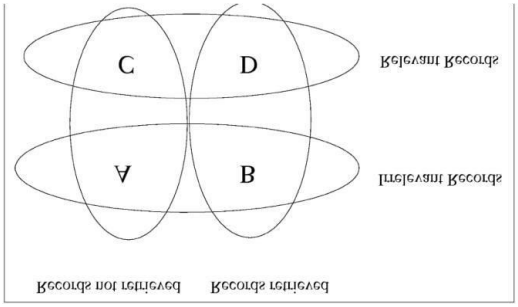
The idea is that such key phrases can be added to the search query to reduce the set of results to those related to the given target individual.

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.