Electronic Sources for Network Analysis

Module 2

Introduction

- Collecting data on social networks required ingenuity from the researcher.
- First, social networks have been studied by observation.
- The disadvantage is the close involvement of the researcher in the process of data collection.
- Standardized surveys minimize the influence of the observer but they rely on an active engagement of the population to be studied.
- Unfortunately achieving a high enough response rate for any survey becomes more and more problematic.

- Data collection using manual methods are extremely ***labor intensive*** and can take up to fifty per cent of the time and resources of a project in network analysis.
- The effort involved in data collection is so immense that network researchers are forced to reanalyze the same data sets over and over in order to be able to contribute to their field.

- A solution to the problem of data collection is to ***reuse*** existing electronic records of social interaction that were not created for the purposes of network analysis.
- ***Scientific communities*** have been studied by relying on publication or project databases showing collaborations among authors or institutes .
- ***Official databases on corporate technology agreements*** allow us to study networks of innovation
- ***Newspaper archives*** are a source of analysis
- These sources often support dynamic studies through historical analysis.

- ***The Internet*** is one data source that is not only vast, diverse and dynamic but also free for all.
- Common to these studies is that they rely entirely on data collected from ***electronic networks*** and ***online information sources***, which allows a complete automation of the data collection process.
- They represent a diversity of social settings and a number of them also exploit the dynamics of electronic data to perform longitudinal analysis.

- Electronic discussion networks
- Blogs and online communities
- Web-based networks

Electronic Discussion Networks

- One of the studies to illustrate the versatility of electronic data is a series of works from the Information Dynamics Labs of Hewlett-Packard.
- Tyler, Wilkinson and Huberman 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.

- They find the study of the email network useful in identifying leadership roles within the organization and finding formal as well as informal communities.
- The authors verify this finding through a set of interviews where they feed back the results to the employees of the Lab

- Adamic and Adar revisits one of the oldest problems of network research, namely the question of ***local search***:
- *how do people find short paths in social networks based on only local information about their immediate contacts?*
- Their findings support earlier results that additional knowledge on contacts such as ***their physical location*** and ***position in the organization*** allows employees to conduct their search much more efficiently.

- The studies of ***electronic communication networks*** based on ***email data*** are limited by privacy concerns.
- Marc Smith and colleagues have published a series of papers on the ***visualization and analysis*** of USENET newsgroups, which predate ***Web-based discussion forums***.
- ***Group communication and collective decision taking*** in various settings are traditionally studied using much more limited written information such as transcripts and records of attendance and voting.

- As in the case with emails Gloor uses the ***headers of messages*** to automatically re-create the discussion networks of the working group.
- The main technical contribution of Gloor is a ***dynamic visualization of the discussion network*** that allows to quickly identify the moments when key discussions take place that activate the entire group and not just a few select members.
- Gloor also performs a ***comparative study across the various groups*** based on the structures that emerge over time.

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 "*personal publishing*" or a "*digital diary*"
- 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, blog rolls and even result in real world meetings such as the Blog Walk series of meetings
- Blogs make a particularly appealing research target due to the ***availability of structured electronic data*** in the form of RSS (Rich Site Summary) feeds.
- RSS feeds contain *the text of the blog posts* and *valuable metadata* such as the timestamp of posts.

- 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

- 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.
- As they are typically used by a much younger demographic they offer an excellent opportunity for studying changes in youth culture

- Most online social networking services (Friendster, Orkut, LinkedIn etc.) closely guard their data even from their own users.
- A technological alternative to these centralized services is the FOAF network.
- FOAF profiles are stored on the web site of the users and linked together using hyperlinks.

Web-based networks

- The content of Web pages is the most endless source of information for social network analysis.
- This content is not only vast, diverse and free to access but also in many cases more up to date than any specialized database.
- On the downside, the quality of information varies significantly and reusing it for network analysis poses significant technical challenges.
- While web content is freely accessible in principle, in practice web mining requires efficient search that at the moment only commercial search engines provide.

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

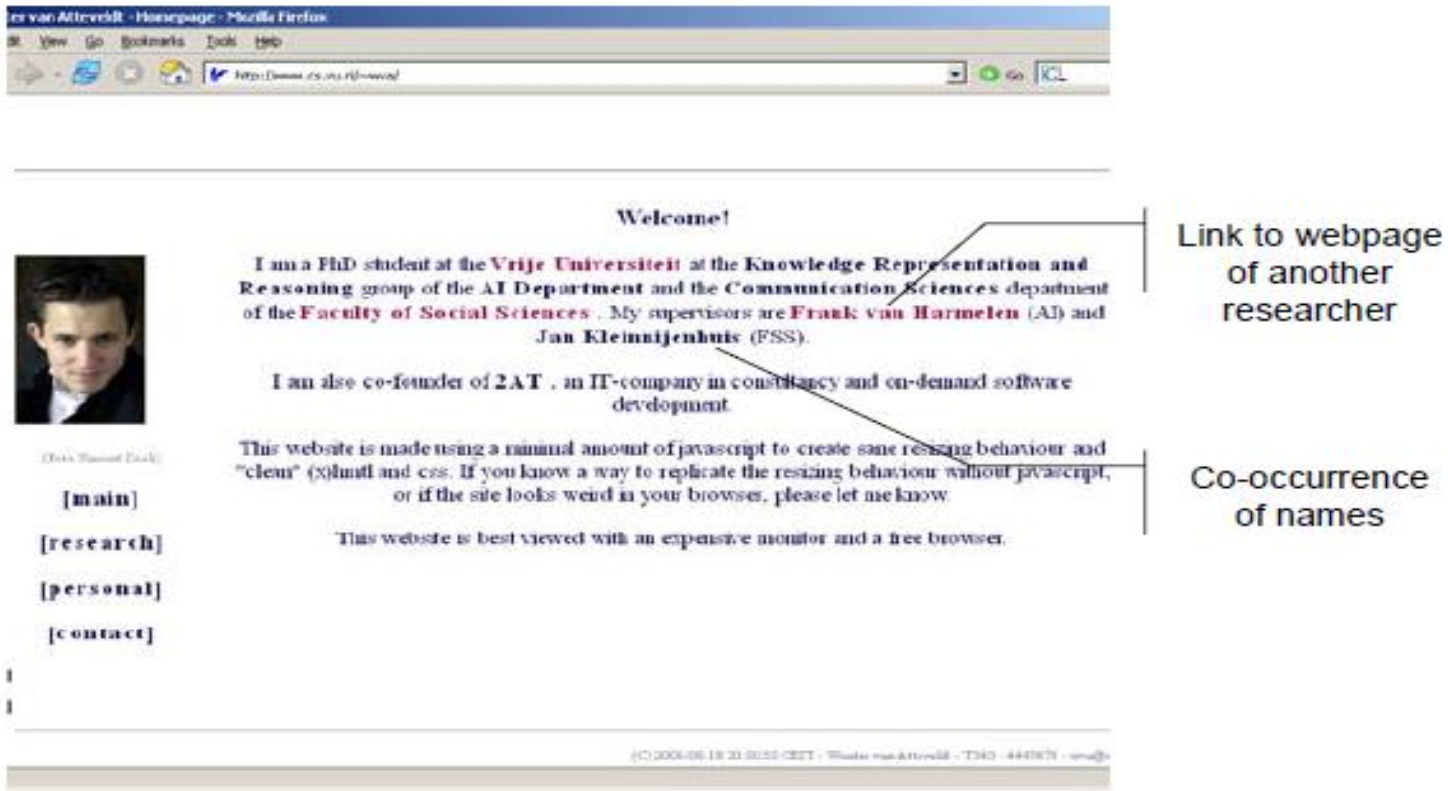

Figure 3.2. Features in web pages that can be used for social network extraction.

- The linking structure of the Web is considered as alternative 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:.
- 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.
- 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 are statistical methods possibly combined with an analysis of the contents of web pages.

- Web mining has been first tested for social network extraction from the Web in the work of Kautz on the ReferralWeb project.
- The goal of Kautz was to build a tool for automating ***referral chaining***: looking for experts with a given expertise who are close to the user of the system, i.e. experts who can be accessed through a chain of referrals.
- The ReferralWeb system extracted connections between the researchers through co-occurrence analysis.
- 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.

- ***Tie strength*** was calculated by dividing the number of co-occurrences with the number of pages returned for the two names individually.
- ***Jaccard-coefficient*** is 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 (only co-occurrences)***.
- If this number has exceeded a certain fixed threshold it was taken as evidence for the existence of a ***tie***.

Figure 3.3. The Jaccard-coefficient is the ratio of the intersection and the union of two sets. In the case of co-occurrence analysis the two sets contain the pages where the individual names occur. The intersection is formed by the pages where both names appear.

- 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: for example, if two names only occur once on the Web and they occur on the same page, their co-efficient will be one.
- However, in this case the absolute sizes are too low to take this as an evidence for a tie.

- A disadvantage of the Jaccard-coefficient is that it penalizes ties between an individual whose name often occurs on the Web and less popular individuals.
  - 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. 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.

- There have been several approaches to deal with ***name ambiguity***.
- Bekkerman and McCallum deal with this problem by using limited background knowledge: 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.
- Bollegala, Matsuo and Ishizuka also apply clustering based on the content similarity.
- 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.

- ***Concept of average precision method***.
- 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).
- 4 types sets are there
  - The records which were retrieved and The records which were not retrieved
  - Relevant records and irrelevant records

- 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 for the computation.
- Let's define *rel(n)* as the relevance at position *n*, where *rel(n)* is *one* if the document at position *n* is the relevant set and *zero* otherwise ($1 \leq n \leq N$).

- Let P(n) denote the precision at position n

$$P(n) = \frac{\sum_{r=1}^{n} rel(r)}{n}$$

Average precision is defined as the average of the precision at all relevant positions:

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

- The ***strength*** is determined by taking the number of the pages where the name of an interest and the name of the person co occur divided by the total number of pages about the person .
- Persons names exhibit the same problems of polysemy and synonymy
- ***Polysemy*** is the association of one word with two or more distinct meaning
- The semantic qualities or sense relations that exist between words with closely related meaning is ***Synonymy***