Social Network Analysis

WEB BASED NETWORK

Course Instructor: Dr.V.S.Felix Enigo

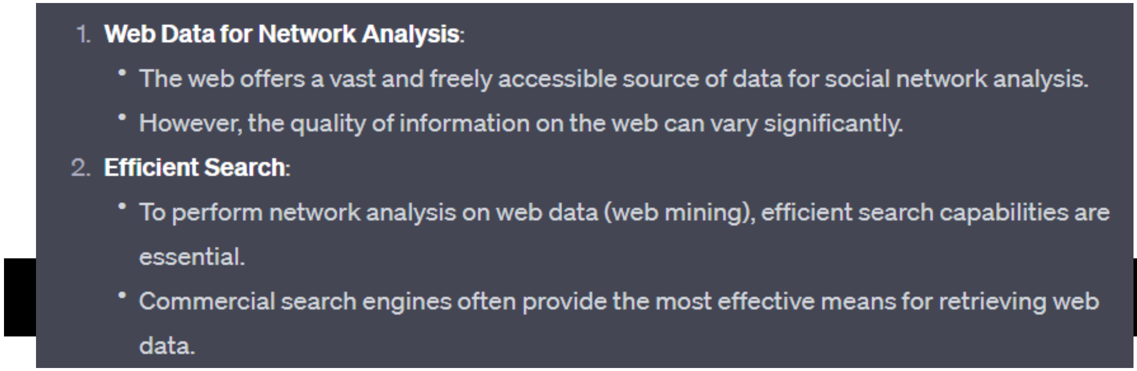
Social Network Analysis

- Web is a vast, diverse and free to access nearly up to date

Downside:

- Quality of information varies significantly

- Reusing for network analysis (web mining) requires efficient search provided by only commercial search engines



1. **Web Data for Network Analysis**:
   - The web offers a vast and freely accessible source of data for social network analysis.
   - However, the quality of information on the web can vary significantly.
2. **Efficient Search**:
   - To perform network analysis on web data (web mining), efficient search capabilities are essential.
   - Commercial search engines often provide the most effective means for retrieving web data.

Social Network Analysis

Two features of web pages considered as basis of extracting social relations:

- Links and co-occurrences are chosen because
  - linking structure represents real world relationships
  - links are authoritative and relevant as it is chosen by author

Social Network Analysis

Drawback :

- Direct links between personal pages are sparse
- Automating searching personal pages for network analysis results in home page search problem
- Linking structure at higher level are studied for network analysis
- Example:
- Heimeriks et al. studied communication and collaboration networks across different fields of research using a multi-layered approach

Social Network Analysis

CO-OCCURENCES

Co-occurrences of names in web pages serve as evidence of relationships

- Extracting relationships based on co-occurrence of names requires web mining
- Requires statistical methods to analyze the contents of web pages
- Web mining first tested for social network extraction by Kautz el al. on ReferralWeb project for *referral chaining*
- *Referral chaining looks for experts with a given expertise close to the user of the system*

Social Network Analysis

- Referralweb extracted through co-occurrence analysis and counts pages using the search engine, AltaVista

- It collected page counts for individual names and number of pages where the names co-occurred

- Disadvantage: very shallow parsing of the web page as indirect references are not counted

Example:

“the president of the United States” will not be associated with George Bush

Social Network Analysis

Jaccard-coefficient (Tie strength) = number of co-occurrences / number of pages returned for the two names individually

- Tie strength ranges 0 – 1
- Jaccard value exceeds certain fixed threshold concluded as a tie
- Jaccard takes relative measure of co-occurrence and not absolute sizes of the sets
- Expertise of individual are extracted using proper name extraction, NLP technique the result is used to extract new names (repeated 2 or 3 times) [snowballing technique]

Social Network Analysis

- Kautz did not evaluated his system for accuracy, but indicated the level of confidence in its decisions
- He proved it is better than official records, as personal pages are more up to date
- Extraction of names and finding tie between names by Search Engine (SE) is a quadratic problem
- Matsuo et al. to reduce the queries for SE first extracted possible contacts from results of SE

Social Network Analysis

- Jaccard-coefficient (JC) penalizes popular ties, but less popular individuals
- To address this, variant of JC is used for confirming a tie
- Variant JC = number of pages for the individual / number of pages for both names

Social Network Analysis

- To compute the strength of association between the name of a given person and a certain topic

$$\text{Tie strength} = \frac{\text{No. of pages found Cooccurences of interest and name of a person}}{\text{Total number of pages about the person}}$$

- Mutschke and Quan Haase, clustered keywords on publications to themes, assign documents to themes, found themes relevant for researcher

Disambiguating Names

- Biggest technical challenge in social network mining is disambiguating person names

- Problem due to polysemy and synonymy

- Polysemy - SE returns partial set of records different variations of name and names with international characters

- Synonymy - Common names return all pages of all names

- Coverage of the Web is very skewed (over-represented) [web pages are largely ranked by popularity]

Social Network Analysis

- Bekkerman and McCallum dealt ambiguity problem using limited background knowledge
- Clustered list of names related to each other, disambiguated based on hyperlinks between the pages, common links or similarity in content

Social Network Analysis

- Weighted directed link between two persons computed as given below:
- Relevant set constitute top ‘n’ pages of ordered list of pages for the first person and a set of pages for the second
- *rel(n), the relevance at position n, is 1 if the relevant document is at position n and zero otherwise (1 ≤ n ≤ N)*

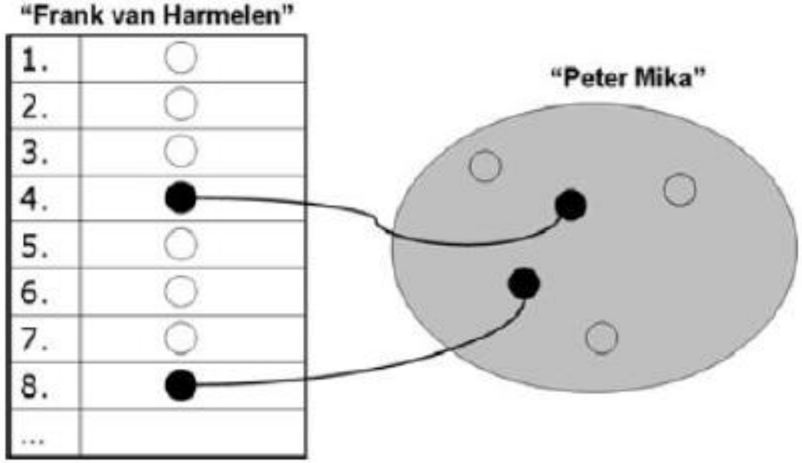
Let *P(n) denote the precision at position n (p@n):*

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

Social Network Analysis

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

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



Summary

- Internet provides vast, free resource of data for analysis
- Links and Co-occurrences are treated as tie between actors
- Direct link between personal pages are too scarce, so indirect methods are sought
- Jaccard Coefficient skewed towards popular ties, so variant used to compute the probability of tie
- Disambiguating names problem is solved by additional knowledge
- Average precision at all position in co-occurences is used as strength of tie between two persons