Electronic Sources of Network Analysis

Data Collection Complexity

- social networks - studied by observation
- Disadvantage - requires close involvement for researcher
- Standardized surveys minimize influence of observer
- Disadvantage - rely on active engagement of population to be studied
- Doubts whether responses are spontaneous and genuine
- Need multiple surveys – to learn network dynamics
- Manual methods labor intensive – 50% time spend on data collection
- Researchers needs to reanalyse to get accurate results
- So need to look for alternatives

Contd…

- Solution - reuse existing electronic records of social interaction
- Examples:
- publication or project databases showing collaborations among authors or institutes
- Corporate DB - study networks of innovation
- News paper archives – study of topics in politics
- But has significant price tag - to access patent databases, media archives, legal and financial records
- Internet - vast, diverse, dynamic and free for all resource
- Rely on electronic networks and online information sources – data collection process automated
- Allows to exploit dynamics of electronic data to perform longitudinal analysis

Electronic Discussion Networks

- Firs study to prove versatility of electronic data - series of works from Information Dynamics Labs of Hewlett-Packard
- Tyler et al – studied communication among employees using corporate email archive
- Recreated network – assigning tie between employee, if exchanged certain threshold of emails
- Also studied leadership role, formal and informal communities
- Adamic and Adar revisits local search – find shortest path using local info
- Than finding most connected people – use of additional info such as physical location and position results quick search
- **Study using email electronic data limits by privacy issue**

Contd…

- Public forums and mailing lists has no privacy issues
- Ex. Analysis of USENET groups
- WWW mailing lists – due to its openness within working group
- Gloor used headers of messages to automatically re-create the discussion networks of working group
- His dynamic visualization quickly identify key discussion

Blogs and Online Communities

- Blogs - “personal publishing” or a “digital diary”
- Analysis of Blogs – used for marketing trend analysis – tools use content anlaysis
- Modern tool enables easy comment , react to comment, bloggers communications – leads to dynamic communities
- This manifests through – aggregated blogs, blog rolls, Blog Walk series of meetings
- Hence blogs can be used to study network analysis

Blogs and online communities

- **Blogs enables researching due to structured data of RSS(Rich Site Summary)**
- RSS aids dynamic analysis – as contains metadata timestamp of the content
- Efimova and Anjewierden - first to study blogs from a communication perspective
- Adar and Adamic offer visualization of communication in blogs

Contd…

- Ex. US election 2004 – exploited blogs to build networks among individual activists and supporters
- Blog analysis – enables marketers to understand interested product choices of young demographics
- Blog Analysis becomes object of study lead to series of Sunbelt social networks conference

Contd…

- Online community spaces and social networking services cater more socialization than blogs e.g. MySpace, LiveJournal
- Used by much younger demographic – enables to study changes in youth culture
- LiveJournal data exposes data such as interest and social networks of users
- Backstrom et al. used this data to study - influence of certain structural properties on community formation and community growth, how changes in the membership of communities relates to change in underlying discussion topics

Contd…

- LiveJournal exposes data for research purposes in a semantic format (an exception, most don’t do)
- Most online social networking services guard the data even from its user e.g. Friendster, Orkut, LinkedIn etc.
- A technological alternative to these centralized services is the FOAF network
- FOAF profiles stored on the website of users, linked together using hyperlinks
- Drawback of FOAF – lack of tools to create, maintain profiles and exploit network

Web based Network

- Web - 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

- Two features of web pages considered as basis of extracting
- social relations:
- links and co-occurrences
  - linking structure represents real world relationships
  - links are authoritative and relevant as it is chosen by author

Contd…

- Drawback :
- Direct links between personal pages are very sparse – as they use browsing of web as mode of navigation (than putting link, update link etc.)
- Automating this task for network analysis, results in home page search problem
- Hence studies are made on linking structure at higher level
- Ex.
- Heimeriks et al. studied communication and collaboration networks across different fields of research using a multi-layered approach

Co-occurences

- Co-occurrences of names in web pages serve as evidence of relationships and frequent phenomenon
- Extracting relationships based on co-occurrence of names requires web mining (as names embedded in text of web)
- This statistical methods is combined with analysis of 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* - looking for experts with a given expertise close to the user of the system

Contd…

- Referralweb extracted through co-occurrence analysis and page counts 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
- Ex. “the president of the United States” will not be associated with George Bush

Jaccard-coefficient

- Tie strength = number of co-occurrences / number of pages returned for the two names individually ***=> Jaccard-coefficient***
- The resulting value is tie strength 0 – 1 zero (no co-occurrences) and one (only co-occurrences)
- If this number exceeds certain fixed threshold was taken as evidence for the existence of a tie
- *Jaccard-coefficient takes* relative measure of co-occurrence and not absolute sizes of the sets
- *Expertise of individual are extracted using proper name extraction,* NLP technique, it is then used to extract new names (repeated 2 or 3 times) [snowballing technique]

Contd…

- Kautz never evaluated his system for accuracy, he proved his recommendation system by, what it is based on and indicate the level of confidence in its decisions

- He proved it is better than official records, as personal pages are more up to date

- Different from Kautz approach, extraction of names and finding tie between names by SE is a quadratic problem, Matsuo et al. first extracted possible contacts from results of search engine for the individual names

- This significantly reduces the number of queries that need to be made to SE at a minimal loss.

Contd…

- Jaccard-coefficient penalizes ties often occurs on the but less popular individuals
- To address this variant is used
- It divides the number of pages for the individual with the number of pages for both names
- if this number reaches a certain threshold, treated as tie

Contd…

- Another approach to calculate the strength of association between the name of a given person and a certain topic

- Strength determined by the number of the pages where the name of an interest and the name of a person co-occur / total number of pages about the person

- Mutschke and Quan Haase, clustered keywords into themes based on the cooccurrences of keywords on publications in bibliography records

- assign documents to themes

- subsequently determines themes relevant for a person based on his or her publications

Problem of ambiguous names

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

- Problem due to polysemy and synonymy

- Different variations of name, names with international characters – SE returns partial set of records

- Common names return all pages of all names

- coverage of the Web can be very skewed (over-represented) [(web pages are largely ranked by popularity]

Methods for disamguiation

- Bekkerman and McCallum deal ambiguity problem using limited background knowledge

- Instead of a single name they assume to have a list of names related to each other

- Disambiguate the appearances by clustering the combined results returned by the search engine (based on hyperlinks between the pages, common links or similarity in content) for the individual names

Contd..

- Bollegala et al went on step further in mining the resulting clusters for key phrases
- i.e. adding key phrases to search query to reduce the set of results to target individual
- For example, when searching for *George Bush the beer brewer one* would add the term *beer to the query*
- *Queries are too specific, results in lower recall*

Average Precision

- When computing weight of directed link between two persons:
- Consider an ordered list of pages for the first person and a set of pages for the second (the relevant set)

- i.e. ask search engine for the top *N* pages for both persons but in the case of the second person the order is irrelevant for computation

- *rel(n), the relevance at position n, where rel(n) is* 1 if the document at position *n is the relevant set and zero otherwise (1 ≤ n ≤ N)*

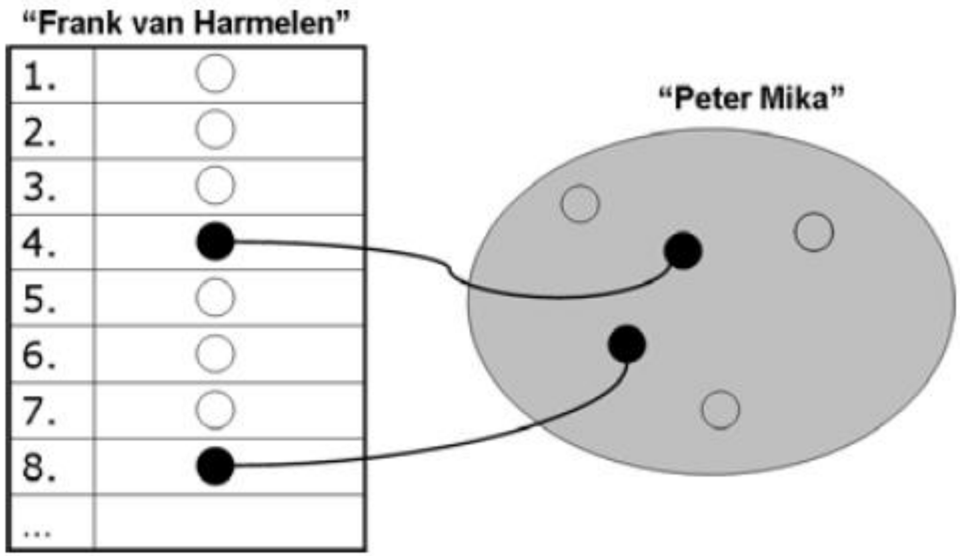
- Let *P(n) denote the precision at position n (p@n):*

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

Average Precision

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

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


"Frank van Harmelen"
1.
2.
3.
4.
5.
6.
7.
8.
...
"Peter Mika"