Social Network Analytics Task Force proposal

TUC meeting

Barcelona, 19-20 Nov.

Why Social Network Analysis? (1)

- Intuitive: everybody knows what a SN is
- SNs are easily represented as a graph
- Scale: from small to large SNs
- Transactional/Analytical
- May integrate different data sources

Why SNA? (2)

- Many different types of uses:
  - Marketing
  - Community management
  - Recommendation
  - Personal use
  - Security
  - Business Intelligence: churn analytics, CRM, etc.
- Many types of queries from the user perspective
  - What can I do with my personal account?
  - What can I recommend you as a user?
  - What are the roles of the users in a SN?
  - Who may be churning in the next week?
  - Who gathers for criminal purposes?
  - How does a customer relate to other customers through my employees?

Types of environments (1)

- Global analytical: need to analyse all the graph
    - Find interesting structural communities/clusters and detect how they are related
    - Evolution of the SN in time to extract patterns
    - Graph metrics: diameter, clustering coefficient…

    Examples:
    1. Create added value links based on patterns or on specific metrics
    2. Find strong communities to recommend the creation of groups
    3. Analyse to predict possible churn from my social network afterwards
    4. Find similar patterns of behaviour

- Partial analytical: part of the graph with possible updates
    - Find patterns of people's behaviour for specific topics: roles, influence, communities
    - Recommendation for topics: products, people

    Examples:
    1. Find the most reputed/influencing people for a specific topic
    2. Find the best researcher in graph databases
    3. Find similar patterns of behaviour for specific keywords

Types of environments (2)

- Transactional
  - Patterns in the behaviour of users
  - Use of personal account, publish posts, add new friends…
  - Specific repeated operations over different parts of the graph
  - Multiple users

  Examples:
  1. I publish a post, then I look for possible friends, etc.
  2. Who are the best friends to send a message with a certain content?

- Integration
  - SN plus taxonomies/ontologies
  - Link data disambiguating the entities

  Examples:
  1. Integrating different bibliographic databases, plus taxonomies using business logic to decide for the
  2. Integrating a merge of two SNs, some users are linked but others may not
  3. Integrating different types of SNs: telephone, SNs, criminal record of individuals

The real situation

- Mixture of on-line transactional & partial analytical
  - I am interacting with the SN
    - Create a new post
    - Find the friends who would like it
    - Search for new possible friends: minimum distance
  - The security/telephone company is traversing the graph to:
    - Find communities based on your characteristics
    - Find specific patterns of behaviour suspicious of churn
  - The Community Manager/on-line support manager/police force:
    - Looks for
- Batch global analytical
- Mixture of batch global analytical & integration
  - I am integrating two SNs, detect similar users
  - Create new added value relationships/links
  - Analyse globally the characteristics

Requirements/graph characteristics

- Delayed commit most of the times may be sufficient
- Data generation: SNs well characterized in literature
- Complex/rich graph
  - Different types of relationships
- On line vs batch
  - Transactional & partial analytical vs global analytical & integration

Choke Points

- Data sets:
  - Do not fit in memory
  - Do not fit in one computer
  - Have to be partitioned across different disks to achieve performance
- Performance:
  - Requires real a certain number of transactions per unit of time
  - Mix priority transactions with lower priority partial analytics
- Algorithmic:
  - Different algorithms to solve the same high level operation
  - Force traversals of the data sets to stress the system

Conclusions

- SN benchmark:
  - Very good match for the needs: intuitive, clear, scalable, etc.
  - Generates a number of workload types
  - Generates a significant number of choke points
  - A lot of industrial partners with SN requirements
  - A lot of knowledge in the community