A comparative study of social network analysis tools

David Combe, Christine Largeron, Előd Egyed-Zsigmond and Mathias Géry

International Workshop on Web Intelligence and Virtual Enterprises 2 (2010)
Outline

Context: social networks and analysis software
Expected functionalities of network analysis software
Benchmark
Conclusion
Definition (Wikipedia)
A social network is a social structure made up of individuals called "nodes," which are tied by one or more specific types of interdependency, such as friendship, common interest, etc.

Sociologic analysis
- Sociological works (Moreno 1934, Milgram 1967, Cartwright and Harary, 1977)
- Web 2.0 : Renewed interest from the Web based social networks websites development.
Context: Social network in business

- For the Gartner Institute:
  - “By 2014, social networking services will replace e-mail as the primary vehicle for interpersonal communications for 20 percent of business users.” (Gartner 2008)
  - Social network analysis is getting mature.

- Some applications in business:
  - Workflow study to adapt management to the real flow in a company;
  - Identify key actors, i.e. for viral marketing.

- These applications need adapted software.
Context: social networks and analysis software

- Network analysis software
  - A previous statistical analysis oriented survey (Huisman & Van Duijn, 2003)

- Networks and needs are changing
  - Size
  - Complex graphs
  - Necessity to make a new benchmark
Context
Expected functionalities of network analysis software
Benchmark
Conclusion
Expected functionalities of network analysis software

1. Representation
2. Visualization
3. Characterization by indicators
4. Community detection
1. Network representation as graph (Cartwright and Harary, 1977)

- Link orientation
  - Undirected links (edges, ex: co-authorship)
  - Directed (arcs, ex: e-mails sent, Enron dataset)

- Weight on edges

- With typed nodes (ex. bipartite network)
1. Network representation as graph

*Vertices 5
*Edges
1 2
1 4
2 3
2 4
3 4
3 5
4 5

Connections

1 → 2, 4
2 → 1, 2, 4
3 → 2, 4, 5
4 → 2, 3, 5
5 → 3, 4

Adjacency list

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Adjacency matrix
2. Visualization

Aim: give a visual representation of the graph, with different approaches:

- Fish eye
  - Centered on an actor
- Force driven visualization layouts
  - Fruchterman Reingold (1984)
    - Iterative algorithm

Random layout  ➔  Force driven visualization  ➔  F-R convergence
3. Characterization by indicators

- Global indicators at network level by:
  - Number of nodes
  - Number of edges
  - Diameter
  - …

- Local indicators at node level:
  - Number of neighbors → degree
  - …

- Distance
  - Length of the shortest path
3. Characterization by indicators: how to decide if an actor is « central »?

- Many ways to determine central actors.
- Ex: Betweenness centrality
  - Which node is the most likely to be an intermediary for a random communication?
  - → higher betweenness centrality
- Selection depends on what they are needed for.
4. Community detection

- **Community**: A set of actors having strong connexions.
- **Community detection algorithms**
  - Walktrap (Latapy & Pons, 2005)
Context
Expected functionalities of network analysis software

Benchmark

Conclusion
Benchmark methodology

- Required points:
  - A social network analysis point of view
  - Scalability
  - Free for educational purposes

- A balance between well established software and newer ones, based on recent development standards (ergonomics, modularity and data portability).

- Datasets: Zachary’s karate-club, DBLP
Software comparison criteria

- Input/output formats
- Custom attribute handling
- Bipartite graphs specific functions
- Longitudinal analysis
- Visualization
- Indicators
- Community detection
Studied software

- **Gephi** is an “interactive visualization and exploration platform”.
- **GUESS** is dedicated to visualization purposes, with several layouts.
- **Tulip** can handle over 1 million vertices and 4 millions edges. It has visualization, clustering and extension by plug-ins capabilities.
- **GraphViz** is mainly for graph visualization.
- **UCI** is not free. It uses Pajek and Netdraw for visualization. It is specialized in statistical and matricial analysis. It calculates indicators (such as triad census, Freeman betweenness) and performs hierarchical clustering.
- **Pajek** is a Windows program for analysis and visualization of large networks. It is freely available, for noncommercial use.
- **igraph** is a free software package for creating and manipulating graphs. It also implements algorithms for some recent network analysis methods.
- **NetworkX** is a package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.
- **JUNG**, for Java Universal Network/Graph Framework, is mainly developed for creating interactive graphs in Java GUIs, JUNG has been extended with some SNA metrics.
Selected software

• Stand-alone software
  ▫ Pajek  http://pajek.imfm.si/doku.php
  ▫ Gephi  http://gephi.org/

• Libraries
  ▫ igraph  http://igraph.sourceforge.net/
  ▫ NetworkX  http://networkx.lanl.gov/
Pajek (Vladimir Batagelj and Andrej Mrvar)

- Development started in 1996
- Data mining oriented
- Many graph operators available
- Fast
- Exports 3D visualization
- Macro
- Supports matrices, adjacency lists and arcs lists oriented input files
Gephi (Bastian M., Heymann S., Jacomy M.)

- Development started in 2008
- Interactive GUI
- Uses Java
- Recent scriptability improvements
- «Photoshop for graphs» with customizable visualization
- Supports the main file formats for networks
- Improvable by plugins
- Community detection still experimental
NetworkX (Brandes U., Erlebach T.)

- Python
- Bipartite graphs ready
- Attribute-friendly
- 1,000,000 nodes wide networks can be handled.
- Lacks in community detection algorithms
- Relies on other software for visualization

```python
>>> import networkx as nx
>>> G=nx.Graph()
>>> G.add_node("spam")
>>> G.add_edge(1,2)
>>> print(G.nodes())
[1, 2, 'spam']
>>> print(G.edges())
[(1, 2)]
>>> G.degree(1)
1
```
Igraph (Csárdi G., Nepusz T.)

- For R (a statistical environment) and Python. The low level routines are written in C.
- GUI available for R.
- Community detection ready.
- Not custom attributes-friendly
How to choose the right tool?

<table>
<thead>
<tr>
<th></th>
<th>Pajek</th>
<th>Gephi</th>
<th>NetworkX</th>
<th>igraph</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input/output</strong></td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Attribute handling</strong></td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>- -</td>
</tr>
<tr>
<td><strong>Bipartite graphs</strong></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Temporality</strong></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td><strong>Visualization</strong></td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>++</td>
</tr>
<tr>
<td><strong>Indicators</strong></td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>+</td>
<td>- -</td>
<td>- -</td>
<td>++</td>
</tr>
</tbody>
</table>

**Mature functionality**

**Not available or weak**
Feature comparison

- Temporality
- Input / output
- Visualization
- Indicators
- Attribute handling

Bipartite Clustering

igraph Pajek NetworkX Gephi
Context
Expected functionalities of network analysis software
Benchmark

Conclusion
Conclusion

- Many domains, many approaches, many software (sociology, computer science, mathematics and physics).
- Functionalities to develop in the future (e.g. for decision support):
  - Temporality awareness
  - Links and nodes attributes analysis
  - Hierarchical graphs
Thank you for your attention.
Any questions?
Bibliography

- Gartner *Hype Cycle for Social Software*, 2008
Bibliography (2)

