State-of-the-art in network data representation

Dr. V.S. Felix Enigo, DCSE, SSNCE

Introduction

- Two fundamental reasons for developing semantic-based representations of social networks:
- Aggregating social network information from heterogeneous environments
- Facilitate the exchange and reuse of case study data in academics of SNA

State of art network data representation

- Most common social data represented as graph with nodes and edges with binary relationship
- Attributes of nodes and edges formalized as function acting on them
- Numerous proprietary formats exists to serialize graphs & attribute data to machine-processable form E.g. Pajek, UCINET

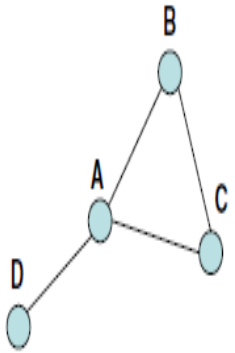
Problems with these approaches

- Pajek and UCINET are incompatible
- Reseacher's first use excel, export to CSV format, yet it is not a graph specific structure
- To be processable by graph packages, additional constraints are put on the content
- Visualization packages use its own proprietary formats

Contd…

- GraphML better in terms of interoperability and extendibility
- GraphML files can be edited, stored, queried, transformed using XML tools
- Graph representations focus on graph structure, a primary input to network analysis and visualization

Graph representations in Pajek, UCINET & GraphML




```
*Vertices 4
1 "A"
2 "B"
3 "C"
4 "D"
*Edges
1 1
1 2
1 3
1 4
2 3
```

```
dl
n = 4
labels embedded
format = edgelist
data:
A B
A C
A D
B C
```

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns">
  <graph id="G" edgedefault="undirected">
    <node id="a"/>
    <node id="b"/>
    <node id="c"/>
    <node id="d"/>
    <edge source="a" target="b"/>
    <edge source="a" target="c"/>
    <edge source="a" target="d"/>
    <edge source="b" target="c"/>
  </graph>
</graphml>
```

Why Semantic based Representation?

- Existing Graph representations stores attribute data separately from network data in Excel sheets, databases or SPSS table

- None of these formats support the aggregation and reuse of electronic data

Why Data Aggregation?

- Example:
- Reuse many data sources describing the same set of their relationships – to analyse and verify same conclusion (also called triangulation)

- For example, apply on email archives, publication databases holding information about researchers

Contd…

- Data sources from multiplex networks use complementary information
- Allow us to study, how these networks differ and how relationships of one type might effect the building of relationships of another type

Requirement for Data Aggregation

- To perform data aggregation across multiple data sources – recognize matching instances in different sources and merge it before analysis

- Graph representations strips social network data (social individuals and their relationships) to nodes and edges required for analysis

Challenges in Data aggregation

- While aggregation, preserving individual identity and relationships crucial to reuse for secondary analysis of data
- Solving these problems requires a very different kind of representation from graph based formats

Semantic-based representation

- A semantic-based representation – exploits the power of ontology languages and tools in aggregating data sets through domain specific knowledge about identity (check if two instances same)

- Enrich our data set with specific domain knowledge (if two people send emails to each other, they know each other - existence of another kind of relationship)

Summary

- The two key problems in aggregating social network data:
- Identification and disambiguation of social individuals
- Aggregation of information about social relationships.