Modelling and Aggregating Social Network Data

Issues in aggregation of Social Network Data

- 1. Maintaining the semantics of social network data is crucial for aggregating social network information
  - Heterogeneous environments –
    - Individual sources of data are under diverse control
- 2. Semantical representations can facilitate the exchange and reuse of case study data in the academic field of Social Network Analysis
  - Data formats not primarily intended for network analysis (excel sheet, SPSS, proprietary graph description languages – that ignore the semantics)

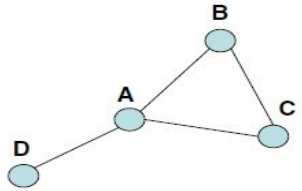
State-of-the-art in network data representation

- Most common kind of social network data can be modelled by a graph where the nodes represent individuals and the edges represent binary social relationships.
- Less commonly, higher-arity relationships may be represented using hyper-edges, i.e. edges connecting multiple nodes.
- Additionally Attributes of nodes and edges - functions operating on it
- Proprietary formats exist to serialize graphs and attribute to machine-processable format
  - Ex. Pajek and UCINET (text-based formats)
  - Issues – both formats incompatible (UCINET has the ability to read and write the .net format of Pajek, but not vice versa.)

Contd...

- Researchers in the social sciences often represent their data initially using Microsoft Excel spreadsheets, exported in the simple CSV (Comma Separated Values) format - not specific to graph structures (it is merely a way to export a table of data)
- Additional constraints on Initial format – as not compatible with graph structure format
- Visualization software packages also have their own proprietary formats (Ex. GraphViz - dot format by AT&T Research)
- Advancement over previos formas – GraphML
  - GraphML defined in XML Schema
  - GraphML interoperable and extensible
  - Processed by XML Tools

Example




```
*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>
```

A simple graph (upper left) described in Pajek .NET, UCINET DL and GraphML formats.

Contd...

- Primary concern - No formats support the aggregation and reuse of electronic data
- Scenario: Need for reusing different data sources for same set of individuals and relationships (need triangulation)
- Triangulation – *use variety of data sources and/or methods* of analysis to verify same conclusion
- Example: email archives and publication databases holding information about researchers

Contd...

- Data sources sometimes contain complementary information
- Example: multiplex network - different kinds of relationships in a community
- In both the cases we need to able to recognize matching instances in different data sources, merge the records and proceed with analysis.
- graph representations strips social network data of exactly to its own characteristics
- Required representation that supports aggregation and reuse (to capture, compare identity of instances and relationships)

Solution

- Semantic-based representation in social networks data (individuals, relationship)
- It uses Ontology languages and Tools through domain specific knowledge about identity
- Example: if two people send emails to each other, they know each other