How can we analyze social networks?

Markus Strohmaier
Denis Helic

Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
dhelic@tugraz.at
web: http://www.kmi.tugraz.at/staff/markus
http://coronet.iicm.edu/denis/homepage
Overview

Today‘s Agenda:

How can we analyze social networks?

A selection of concepts from Social Network Analysis

• Sociometry, adjacency lists and matrices
• Affiliation networks
• KNC Plots
• Prominence
• Cliques, clans and clubs
Sociometry as a precursor of (social) network analysis
[Wasserman Faust 1994]

• Jacob L. Moreno, 1889 - 1974
• Psychiatrist

• born in Bukarest, grew up in Vienna, lived in the US
• Worked for Austrian Government

• Driving research motivation (in the 1930‘s and 1940‘s):
  – Exploring the advantages of picturing interpersonal interactions using sociograms, for sets with many actors
Sociometry
[Wassermann and Faust 1994]

- Sociometry is the study of positive and negative relations, such as liking/disliking and friends/enemies among a set of people.

- A social network data set consisting of people and measured affective relations between people is often referred to as *sociometric*.

- Relational data is often presented in two-way matrices termed *sociomatrices*.

Can you give an example of web formats that capture such relationships?

Sociometry
[Wassermann and Faust 1994]

- Images taken from Wasserman/Faust page 76 & 82

Table 3.1. Sociomatries for the six actors and three relations of Figure 3.2

<table>
<thead>
<tr>
<th>Allison</th>
<th>Drew</th>
<th>Elliot</th>
<th>Keith</th>
<th>Ross</th>
<th>Sarah</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Friendship at Beginning of Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allison</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Drew</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Elliot</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Keith</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Ross</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Sarah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Friendship at End of Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allison</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Drew</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Elliot</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Keith</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Ross</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Sarah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Lives Near</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allison</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Drew</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Elliot</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Keith</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Ross</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Sarah</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 3.2. The six actors and the three sets of directed lines — a multivariate directed graph.
Fundamental Concepts in SNA
[Wassermann and Faust 1994]

- **Actor**
  - Social entities
  - Def: Discrete individual, corporate or collective social units
  - Examples: people, departments, agencies

- **Relational Tie**
  - Social ties
  - Examples: Evaluation of one person by another, transfer of resources, association, behavioral interaction, formal relations, biological relationships

- **Dyad**
  - Emphasizes on a tie between two actors
  - Def: A dyad consists of two actors and a tie between them
  - An inherent property between two actors (not pertaining to a single one)
  - Analysis focuses on dyadic properties
  - Example: Reciprocity, trust

Which networks would not qualify as social networks?
Fundamental Concepts in SNA
[Wassermann and Faust 1994]

• Triad
  – Def: A subgroup of three actors and the possible ties among them

  ![Diagram of four possible triadic states in a graph](image)

  **Transitivity**
  – If actor i „likes“ j, and j „likes“ k, then i also „likes“ k

  **Balance**
  – If actor i and j like each other, they should be similar in their evaluation of some k
  – If actor i and j dislike each other, they should evaluate k differently

Example 1: Transitivity
Example 2: Balance
Example 3: Balance
Fundamental Concepts in SNA
[Wassermann and Faust 1994]

• **Definition of a Social Network**
  - Consists of a finite set or sets of actors and the relation or relations defined on them
  - Focuses on *relational* information rather than attributes of actors
One and Two Mode Networks

[Wasserman Faust 1994]

• The mode of a network is the number of sets of entities on which structural variables are measured.

• The number of modes refers to the number of distinct kinds of social entities in a network.

• One-mode networks study just a single set of actors.

• Two mode networks focus on two sets of actors, or on one set of actors and one set of events.
Affiliation Networks

- Affiliation networks are two-mode networks
  - Nodes of one type “affiliate” with nodes of the other type (only!)
- Affiliation networks consist of subsets of actors, rather than simply pairs of actors
- Connections among members of one of the modes are based on linkages established through the second
- Affiliation networks allow to study the dual perspectives of the actors and the events

![Bipartite and Complete Bipartite Graphs](Fig. 4.15. Bipartite graphs [Wasserman Faust 1994])
Is this an Affiliation Network? Why/Why not?

FIG. 8 Friendship network of children in a US school. Friendships are determined by asking the participants, and hence are directed, since A may say that B is their friend but not vice versa. Vertices are color coded according to race, as marked, and the split from left to right in the figure is clearly primarily along lines of race. The split from top to bottom is between middle school and high school, i.e., between younger and older children. Picture courtesy of James Moody.
Examples of Affiliation Networks on the Web

• Facebook.com users and groups/networks
• XING.com users and groups
• Del.icio.us users and URLs
• Bibsonomy.org users and literature
• Netflix customers and movies
• Amazon customers and books
• Scientific network of authors and articles
• etc
Representing Affiliation Networks
As Two Mode Sociomatrices

[Wasserman Faust 1994]

General form:

\[
\begin{pmatrix}
0 & A \\
A' & 0
\end{pmatrix}
\]

Fig. 8.3. Sociomatrix for the bipartite graph of six children and three parties
Two Mode Networks and One Mode Networks

- **Folding** is the process of transforming two mode networks into one mode networks
  - Also referred to as: $T, \perp$ projections [Latapy et al 2006]
- Each two mode network can be folded into 2 one mode networks

### Type A

- A
- B
- C

### Type B

- I
- II
- III
- IV

#### Examples:
- Two mode network: conferences, courses, movies, articles
- 2 One mode networks: actors, scientists, students
Transforming Two Mode Networks into One Mode Networks

[Wasserman Faust 1994]

• Two one mode (or co-affiliation) networks (folded from the children/party affiliation network)

\[ M_P = M_{PC} \times M_{PC}' \]

\[
\begin{array}{ccccccc}
  & n_1 & n_2 & n_3 & n_4 & n_5 & n_6 \\
n_1 & 2 & 0 & 1 & 1 & 2 & 1 \\
n_2 & 0 & 1 & 1 & 0 & 1 & 1 \\
n_3 & 1 & 1 & 2 & 1 & 2 & 1 \\
n_4 & 1 & 0 & 1 & 1 & 1 & 0 \\
n_5 & 2 & 1 & 2 & 1 & 3 & 2 \\
n_6 & 1 & 1 & 1 & 0 & 2 & 2 \\
\end{array}
\]

Fig. 8.5. Actor co-membership matrix for the six children

\[
\begin{array}{ccc}
m_1 & m_2 & m_3 \\
m_1 & 3 & 2 & 2 \\
m_2 & 2 & 4 & 2 \\
m_3 & 2 & 2 & 4 \\
\end{array}
\]

Fig. 8.6. Event overlap matrix for the three parties

[Images taken from Wasserman Faust 1994]
Transforming Two Mode Networks into One Mode Networks

[Wasserman Faust 1994]

\[ M_P = M_{PC} \times M_{PC}' \]

C...Children
P...Party

\[
\begin{array}{cccccc}
\text{Allison} & \text{Drew} & \text{Eliot} & \text{Keith} & \text{Ross} & \text{Sarah} \\
\hline
\text{Party 1} & 1 & 0 & 0 & 0 & 1 & 0 \\
\text{Party 2} & 0 & 1 & 1 & 0 & 1 & 1 \\
\text{Party 3} & 1 & 0 & 1 & 1 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{Allison} & \text{Drew} & \text{Eliot} & \text{Keith} & \text{Ross} & \text{Sarah} \\
\hline
\text{Party 1} & 1 & 0 & 0 & 0 & 1 & 0 \\
\text{Party 2} & 0 & 1 & 1 & 0 & 1 & 1 \\
\text{Party 3} & 1 & 0 & 1 & 1 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{Allison} & \text{Drew} & \text{Eliot} & \text{Keith} & \text{Ross} & \text{Sarah} \\
\hline
\text{Party 1} & 1 & 0 & 0 & 0 & 1 & 0 \\
\text{Party 2} & 0 & 1 & 1 & 0 & 1 & 1 \\
\text{Party 3} & 1 & 0 & 1 & 1 & 1 & 0 \\
\end{array}
\]

Output: Weighted regular graph
Transforming Two Mode Networks into One Mode Networks

[Wasserman Faust 1994]
## Set-theoretic/Vector-based Measures of Similarity


### Similarity between P1 & P2:

**Raw measure** (or *Simple matching coefficient, result of folding*)

\[ |X \cap Y| = 2 \]

(does not take into account sizes of X or Y)

**Binary Approaches (incl. Normalization)**

- **Dice’s coefficient (D)**
  \[ 2 \frac{|X \cap Y|}{|X| + |Y|} = \frac{2}{7} \]

  \[ 2*2/(3+4) = 4/7 \]

- **Jaccard’s coefficient (J)**
  \[ \frac{|X \cap Y|}{|X \cup Y|} = \frac{2}{5} \]

- **Cosine coefficient (C)**
  \[ 2/(|X|^{1/2} \times |Y|^{1/2}) \approx 0.577 \]

- **Overlap coefficient (O)**
  \[ \frac{|X \cap Y|}{\min(|X|, |Y|)} = \frac{2}{3} \]

All the left (except the raw measure) are normalized similarity measures:

1. For \( S = D, J, C, O \), \( S(X,Y) = S(Y,X) \) and \( S(X; Y) = 1 \) iff \( X = Y \).
2. For \( S = D, J, C, O \), \( 0 \leq S(X,Y) \leq 1 \)


### Vector interpretation (P1, P2)

<table>
<thead>
<tr>
<th>Party 1</th>
<th>Party 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Allison  
Drew  
Eliot  
Keith  
Ross  
Sarah

Counting measure \(| . |\) gives the size of the set.

---

[cf. http://www.dcs.gla.ac.uk/Keith/Chapter.3/Ch.3.html]
# Real-valued Vectors

<table>
<thead>
<tr>
<th></th>
<th>Binäre Vektoren ($^1$)</th>
<th>Vektoren mit reellen Werten ($^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Measure</strong></td>
<td>$</td>
<td>X \cap Y</td>
</tr>
<tr>
<td><strong>Dice-Coefficient</strong></td>
<td>$\frac{2</td>
<td>X \cap Y</td>
</tr>
<tr>
<td><strong>Jaccard-Coefficient</strong></td>
<td>$\frac{</td>
<td>X \cap Y</td>
</tr>
<tr>
<td><strong>Cosine-Coefficient</strong></td>
<td>$\frac{</td>
<td>X \cap Y</td>
</tr>
<tr>
<td><strong>Overlap-Coefficient</strong></td>
<td>$\frac{</td>
<td>X \cap Y</td>
</tr>
</tbody>
</table>

\[ |\mathbf{\hat{x}}| = \sqrt{\sum_{i=1}^{n} \hat{x}_i^2} \]

\[ \mathbf{\hat{x}} \cdot \mathbf{\hat{y}} = \sum_{i=1}^{n} x_i y_i \]

---

(C) Karin Haenelt

nis Helic 2011

$^1$(Manning/Schütze, 2000, 300/301)

$^2$(Ferber, 2003)
The k-neighborhood graph, $G_k$

Given bipartite graph B, users on left, interests on right

Connect two users if they share at least $k$ interests in common

The $k$-neighborhood graph, $G_k$

Given bipartite graph $B$, users on left, interests on right

Connect two users if they share at least $k$ interests in common
The k-neighborhood graph, $G_k$

Given bipartite graph B, users on left, interests on right

Connect two users if they share at least $k$ interests in common

The $k$-neighborhood graph, $G_k$

Given bipartite graph $B$, users on left, interests on right

Connect two users if they share at least $k$ interests in common

Illustration $k=1$

Illustration $k=2$

Illustration $k=3$

Illustration $k=4$

Illustration $k=5$

The KNC-plot

The k-neighbor connectivity plot

- How many connected components does $G_k$ have?
- What is the size of the largest component?

Answers the question:

how many shared interests are meaningful?

- Communities, Cuts
Analysis

Four graphs:

- **LiveJournal**
  - Blogging site, users can specify interests
- **Y! query logs** *(interests = queries)*
  - Queries issued for Yahoo! Search *(Try it at www.yahoo.com)*
- **Content match** *(users = web pages, interests = ads)*
  - Ads shown on web pages
- **Flickr photo tags** *(users = photos, interests = tags)*

All data anonymized, sanitized, downsampled

- Graphs have 100s of thousands to a million users
Examples

- Largest component
- Number of components

At $k=5$, all connected.
At $k=6$, interesting!

Content match
Web pages = “users”
Ads = “interests”

At $k=6$, nobody connected

Flickr
Photos = “users”
Tags = “interests”

A node, $n_i$, is a cutpoint if the number of components in a graph $G$ that contains $n_i$ is fewer than the number of components in the subgraph that results from deleting $n_i$ from the graph.

Cutpoint or „Articulation point“

Analogous to the concept of bridges, Wasserman p113

Which node(s) represents a cutpoint? Why?
The Web Graph is Flat

Book tip
„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)
http://www.geom.uiuc.edu/~banchoff/Flatland/

How can we infer information about the nth+1 dimension?

E.g. popularity, trust, prestige, importance, …

http://www.youtube.com/watch?v=BWyTxCsIXE4
http://www.flatlandthefilm.com/
Inhabitants of Flatland

Tradesman

Men (The hero in this novel is A. Square)

Woman

Priests

Book tip
„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)
http://www.geom.uiuc.edu/~banchoff/Flatland/
Recognition by sight

Book tip

„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)
http://www.geom.uiuc.edu/~banchoff/Flatland/
What kind of information can we infer from a „flat“ social graph?
Centrality and Prestige
[Wasserman Faust 1994]

Which actors are the most important or the most prominent in a given social network?

What kind of measures could we use to answer this (or similar questions)?

What are the implications of directed/undirected social graphs on calculating prominence?

⇒ In directed graphs, we can use Centrality and Prestige
⇒ In undirected graphs, we can only use Centrality
Prominence
[Wasserman Faust 1994]

We will consider an actor to be prominent if the ties of the actor make the actor particularly visible to the other actors in the network.
Actor Centrality
[Wasserman Faust 1994]

Prominent actors are those that are extensively involved in relationships with other actors.

This involvement makes them more visible to the others.

No focus on directionality -> what is emphasized is that the actor is involved.

A *central actor* is one that is involved in many ties. [cf. Degree of nodes]
Actor Prestige
[Wasserman Faust 1994]

A prestigious actor is an actor who is the object of extensive ties, thus focusing solely on the actor as a recipient.

[cf. indegree of nodes]

Only quantifiable for directed social graphs.

Also known as *status, rank, popularity*
Different Types of Centrality in Undirected Social Graphs
[Wasserman Faust 1994, Scripps et al 2007]

Degree Centrality

- **Actor Degree Centrality:** Based on degree only

Closeness Centrality

- **Actor Closeness Centrality:**
  - Based on how close an actor is to all the other actors in the set of actors
  - Closeness is the reciprocal of the sum of all the geodesic (shortest) distances from a given node to all others
  - Nodes with a small CC score are closer to the center of the network while those with higher scores are closer to the edge.

Betweenness Centrality

- **Actor Betweenness Centrality:**
  - An actor is central if it lies between other actors on their geodesics
  - The central actor must be between many of the actors via their geodesics

⇒ All three can be normalized to a value between 0 and 1 by dividing it with its max. value

\[
\tilde{C}_D(n_i) = \sum_j I[(i, j) \in E]
\]

Where \(I\) is a 0-1 indicator function.

\[
C_C(n_i) = \left[ \sum_{j=1}^N d(n_i, n_j) \right]^{-1}
\]

\(d(u; v)\) is the geodesic distance from \(u\) to \(v\).

\[
C_B(n_i) = \sum_{j < k} \frac{g_{jk}(n_i)}{g_{jk}}
\]

where \(g_{jk}\) is the number of geodesic paths from \(j\) to \(k\) \((j, k \text{ all pairs of nodes})\) and \(g_{jk}(ni)\) is the number of geodesic paths from \(j\) to \(k\) that go through \(i\).
Centrality and Prestige in Undirected Social Graphs
[Wasserman Faust 1994]

Actor = closeness
= betweenness
centrality:
n1>n2,n3,n4,n5,n6
,n7

Actor centrality =
Betweenness centrality
= Closeness centrality:
n1=n2=n3=n4=n5=n6
=n7

Betweenness
centrality:
n1>n2,n3>n4,n5>n
6,n7

Fig. 5.1. Three illustrative networks for the study of centrality and
prestige
Centrality and Prestige in Undirected Social Graphs
[Wasserman Faust 1994]

Examples:
http://www.faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html
How can we identify groups and subgroups in a social graph?
Cliques, Subgroups
[Wasserman Faust 1994]

Definition of a Clique
• A clique in a graph is a maximal complete subgraph of three or more nodes.

Remark:
• Restriction to at least three nodes ensures that dyads are not considered to be cliques
• Definition allows cliques to overlap

Informally:
• A collection of actors in which each actor is adjacent to the other members of the clique

Fig. 7.1. A graph and its cliques
Subgroups
[Wasserman Faust 1994]

Cliques are very strict measures

• Absence of a single tie results in the subgroup not being a clique
• Within a clique, all actors are theoretically identical (no internal differentiation)
• Cliques are seldom useful in the analysis of actual social network data because definition is overly strict

⇒ So how can the notion of cliques be extended to make the resulting subgroups more substantively and theoretically interesting?

⇒ Subgroups based on reachability and diameter
n cliques
[Wasserman Faust 1994]

N-cliques require that the geodesic distances among members of a subgroup are small by defining a cutoff value $n$ as the maximum length of geodesics connecting pairs of actors within the cohesive subgroup.

An n-clique is a maximal complete subgraph in which the largest geodesic distance between any two nodes is no greater than $n$.

NOTE: Geodesic distance between 4 and 5 "goes through" 6, a node which is not part of the 2-clique.

Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs.
n clans
[Wasserman Faust 1994]

An n-clan is an n-clique in which the geodesic distance between all nodes in the subgraph is no greater than n for paths within the subgraph.

N-clans in a graph are those n-cliques that have diameter less than or equal to n (within the graph).

⇒ All n-clans are n-cliques.

Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs

Which 2-clans can you identify in the following graph?

Why is \{1,2,3,4\} not a 2-clan?

Why is \{1,2,3,4,5\} not a 2-clan?
n clubs

[Wasserman Faust 1994]

An n-club is defined as a maximal subgraph of diameter n.

A subgraph in which the distance between all nodes within the subgraph is less than or equal to n and no nodes can be added that also have geodesic distance n or less from all members of the subgraph.

⇒ All n-clubs are contained within n-cliques.
⇒ All n-clans are also n-clubs
⇒ Not all n-clubs are n-clans

Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs
Subgroups in Co-Affiliation Networks

Borgatti 1997

• The obvious next step would be to try to identify these subgroups in co-affiliation networks.
  – For example, we can search for cliques, n-cliques, n-clans, n-clubs.

• Unfortunately, these methods are not well suited for analysing a bipartite graph.
  – In fact, bipartite graphs contain no cliques
  – In contrast, bipartite graphs contain too many 2-cliques and 2-clans.
  – One of the problems is that, in the bipartite graph, all nodes of the same type are necessarily two links distant.

⇒ we need to consider special types of subgraphs which are more appropriate for two-mode data.
Bicliques
[Borgatti 1997]

A biclique is a maximal complete bipartite subgraph of a given bipartite graph.

Reasonable to insist on bicliques of the form $K_{m,n}$ where $m$ and $n$ are greater than 2

- Why? Each of the two modes should form (after folding) interesting structures (triads or greater)
Subgroups in Co-Affiliation Networks
Borgatti 1997

- Clearly, we can define extensions of n-cliques, n-clubs and n-clans to n-bicliques, n-biclubs and n-biclans.
- But, the extensions would in many senses be unnatural since n would need to be odd.

- Next week we will discuss a way to analyze subgroups in affiliation networks: Galois Lattices
Any questions?