Social Network Analysis

Anuška Ferligoj

Faculty of Social Sciences, University of Ljubljana, Slovenia

Discovering Science 2014
Bled, Slovenia, 8-10 October, 2014

October 8, 2014

Outline

Introduction

Social Network Analysis (SNA) has attracted considerable interest from social and behavioral science community in recent decades.

Much of this interest can be attributed to the focus of social network analysis on *relationship* among units, and on the patterns of these relationships.

Social network analysis is a rapidly expanding and changing field with broad range of approaches, methods, models and substantive applications.

Basic definitions

Social network consists of *units* (social actors) with one or more social *relations* (set of ties or relationships) defined over those units. *Properties* on units and on ties can be added.

Units can be individuals, groups, organizations, ...

A relation among workers in an organization can be 'to communicate with'.

The relation can be *symmetric* (e.g., co-authorship) or *non-symmetric* (e.g., friendship).

Some applications of SNA

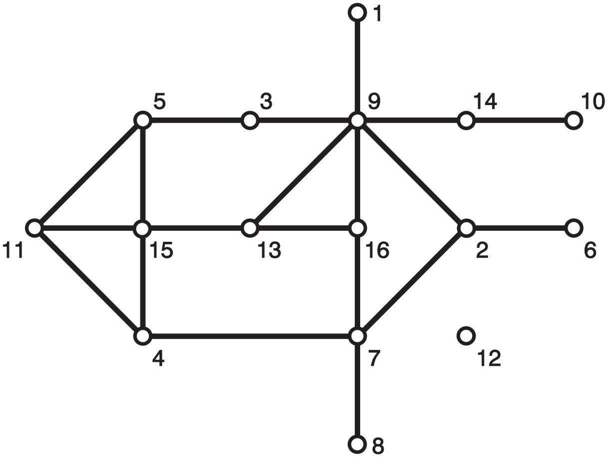
- personal communities
- social support
- animal social networks
- networking online
- corporate elits and intercorporate networks
- policy networks
- social movements and collective action
- crime
- terrorist networks
- scientific and scholarly networks
- cultural networks
- spacial networks
- intra- and inter-organizational studies
- health and illness, particularly AIDS
- world political and economic system

Presentation of a social network

A social network can be **presented** by

- graph
- relational matrix
- set of related pairs of units

Example: Graph



Relational matrix

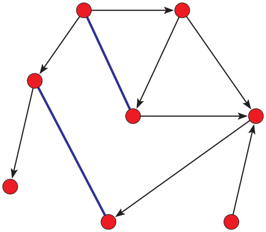
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
5	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
14	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
15	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0
16	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0

Set of related pairs of units

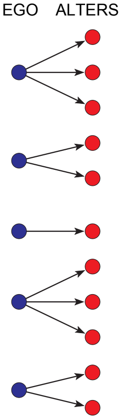
$$R = \{(1:9),(2:6),(2:7),(2:9),(3:5),(3:9),(4:7),(4:11),$$
$$(4:15),(5:11),(5:15),(7:8),(7:16),(9:13),$$
$$(9:14),(9:16),(10:14),(11:15),(13:15),(13:16)\}$$

Whole and egocentered networks

- When the ties for each pair of units are known, we have a *whole network*.
- If a set of units are given (e.g., a random sample) and only ties from each of these units (*egos*) to some (other) units (*alters*) are measured (usually not ties between these alters) we speak about *egocentered networks* or *personal networks*.



WHOLE NETWORK



EGO-CENTERED NETWORKS

Types of networks

Besides the usual whole networks some extended types of networks are also used:

- *2-mode networks*,
- *signed networks*
- *temporal networks* or dynamic networks,
- *multiple networks* or multi-relational networks,
- *specialized networks* (e.g., genealogies).

Two-mode networks

One-mode social network is defined by the set of units and the relationships defined only between them, e.g., friendship relation among pupils in a class.

If there are two sets of units and the relationships are defined between units of the first set and units of the second set we referred to the *two-mode social network*.

An example of two-mode network is membership network, e.g., scientists (first set) are members of different scientific associations (second set).

Deep South

Classical example of two-mode network are the Southern women (Davis 1941).

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

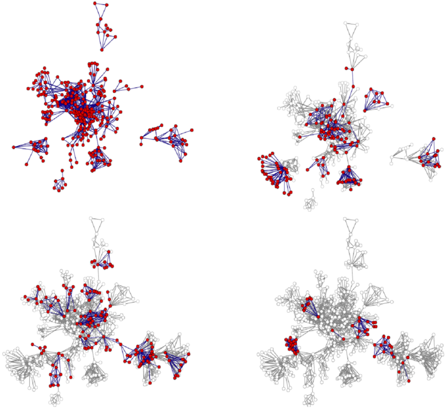
Signed networks

Here, the ties can have positive, zero or negative values.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	-2	3	0	0	0	-3	0	0	-1	0	1	0	2	0	0	0	0
2	3	0	0	-3	0	0	1	-2	0	0	0	2	-1	0	0	0	0	0
3	3	-2	0	-3	0	-2	0	0	0	0	0	0	2	0	0	-1	1	2
4	-2	-3	0	0	3	1	0	0	0	0	2	0	0	-1	0	0	0	0
5	0	0	0	3	0	0	0	0	1	0	2	0	0	0	0	0	0	0
6	0	-1	-3	3	1	0	-2	0	2	0	0	0	0	0	0	0	-2	0
7	0	3	0	-3	0	-2	0	-1	0	0	0	1	0	0	0	2	0	0
8	0	-3	-2	3	0	2	0	0	1	0	0	0	0	-1	0	0	0	0
9	0	0	-3	0	1	0	0	3	0	0	0	2	0	0	0	0	-2	-1
10	0	0	0	3	1	0	0	0	1	0	0	0	2	0	0	0	0	0
11	-1	-3	-2	0	2	0	0	3	0	0	0	0	0	1	0	0	0	0
12	3	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
13	0	-3	0	0	2	-2	1	0	0	0	0	-1	0	0	0	0	0	3
14	3	0	0	-3	0	0	0	-2	0	0	0	1	0	0	2	0	-1	0
15	0	3	-2	-1	0	0	1	0	0	0	0	2	-3	0	0	0	0	0
16	0	3	-1	-3	0	0	2	0	0	0	0	0	0	0	1	0	-2	0
17	0	1	2	-1	0	-3	0	-2	0	0	0	0	0	0	0	0	0	3
18	0	1	2	-1	0	0	0	-3	0	-2	0	0	0	0	0	0	3	0

Temporal networks

Temporal or *dynamic* networks change over time. Here, units or ties can change through time.



Moody:
Drug users in Colorado Springs, 5 years

Multiple networks

If several relations are defined over the set of units, such a network is called *multi-relational* network or *multiple* network.

Example

Sampson (1968) reported data about four relations (positive and negative ties) at five time points among a group of 18 trainee monks at a New England Monastery:

- affect,
- esteem,
- influence, and
- sanctioning.

Therefore, it is multiple and temporal signed network.

Unit of the observation

Unit of the observation and analysis can be:

- a social actor
- a dyad
- a triad
- a subgroup
- a network

Important researchers in SNA

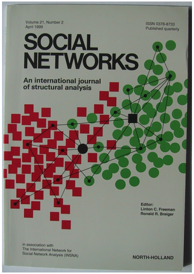
Graph theory: Euler (1736), Hamilton (cycles), Kirchoff (electrical circuits), Kekule (molecule), Kruskal, Dijkstra, Ford and Fulkerson (optimization), Harary, Berge (first books) . . .

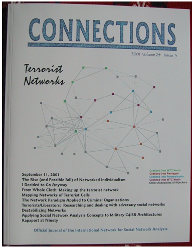
Moreno

- Moreno (1934) – sociometry
- Small networks: Lewin (1936), Warner and Lunt (1941), Heider (1946), Bavelas (1948) – centrality, Homans (1950), Cartwright and Harary (1956), Mitchell (1969)
- Roles and positions: Nadel (1957), H. White (1970)
- Probabilistic models: Wasserman (1977), Leinhardt and Holland (1981), Frank (1986)

Freeman L.C. (2004) The Development of Social Network Analysis

Institutionalized forum of SNA





- International Association of Social Network Analysis – INSNA, 1978
- Journal: Social Networks, 1978
- Newsletter: Connections, 1978
- SUNBELT conferences, 1981
- e-Journal: Journal of Social Structure, 2000
- Journal: Social Network Analysis and Mining, 2011
- Journal: Network Science, 2013

Network data collection

- Archival records
- Observation
- Informant data
- Diary
- Survey
- Network data collection from Internet and data bases
- Other data collection techniques

An example of archival network data

Padgett collected the network and attribute data in the very rich archives in Florence for the most important 116 Florentine families in 15th century.

His reasearch question was: Why the Medici family got the power in Florence in the 15th century (1434)?

He collected the following attribute data:

- the family wealth (measured in the year 1427) and
- the number of council seats held by family members in the years 1282-1344.

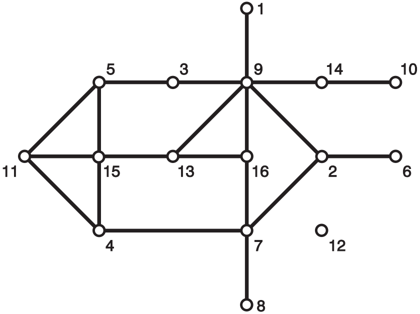
Attribute data

		family wealth	council seats
Acciaiuoli	1	10.448	53
Albizzi	2	35.730	65
Barbadori	3	55.351	N/A
Bischeri	4	44.378	12
Castellani	5	19.691	22
Ginori	6	32.013	N/A
Guadagni	7	8.127	21
Lamberteschi	8	41.727	0
Medici	9	103.140	53
Pazzi	10	48.233	a
Peruzzi	11	49.313	42
Pucci	12	2.970	0
Ridolfi	13	26.806	38
Salviati	14	9.899	35
Strozzi	15	145.896	74
Tornabuoni	16	48.258	N/A

N/A indicates "not available data"
a indicates a special case of Pazzi family

Marriage ties among 16 Florentine families

The Medici family had not the highest economic nor political power. Why they became the leading family in Florence? Let us look on the marriage ties of these families:



1. Acciaiuoli
2. Albizzi
3. Barbadori
4. Bischeri
5. Castellani
6. Ginori
7. Guadagni
8. Lamberteschi
9. Medici
10. Pazzi
11. Peruzzi
12. Pucci
13. Ridolfi
14. Salviati
15. Strozzi
16. Tornabuoni

An example of observational data

It was mentioned before that Sampson (1968) reported data about four signed relations at five time points among a group of 18 trainee monks at a New England Monastery. He used a qualitative approach to observe the relationships among the monks and later constructed the relational matrices.

Sampson collected data for four relations:

- affect,
- esteem,
- influence, and
- sanctioning.

Survey network data collection

Surveys are widely used to collect data on ties among units. Surveys remain vital source of network data for many situations in which direct observation, diaries and other methods of collecting network data are impractical.

In survey data collection we have to consider the following main dilemas:

- which *mode* to use (face-to-face interview, telephone interview, mail questionnaire, web questionnaire,...);
- *free* or *fixed choices* in naming the related units;
- *recognition* (complete listing or roster available) or *free recall*.

Instruments for network data

Different approaches commonly used in standardized questionnaires and interviews to obtain data on social networks are used for measuring

- *whole networks* and
- *egocentric networks*; here, we distinguish between
    - 'name generator' to elicit a list of alters within a respondent's (ego's) egocentric network,
    - 'name interpreter' to obtain information about alters and ties between ego and alters.

More about survey instruments to measure social networks is in Marsden (2011).

Important Issues

- Network boundary specification
- Quality assessment of network measures (e.g., reliability, validity)

Network boundary specification

Deciding on the set(s) of units that lie within a network is a difficult problem for network studies.

Boundary specification strategies (Marsden, 2011):

- *positional approach* based on characteristics of units or formal membership criteria (e.g., employment by an organization, assignment to a school classroom),
- *event-based approach* resting on participation in some class of activities (e.g., participants of a selected event in a time interval),
- *relational approach* based on social connectedness (e.g., studies of service delivery systems where some core agencies are defined and later added others to which they refere as clients).

Network data quality

Unfortunately, we cannot measure without measurement error. Key questions include:

- How much error is there in a certain measurement?
- What is the quality of the resulting measurements from using an instrument?
- Which measurement instrument produces better measurements?

As measurement errors can effect the structure of a network significantly, the effect of question wording and methods of naming related units on the results have to be studied more systematically also in the field of social network analysis.

The quality of measurement can be estimated by reliability and validity.

Estimating reliability

Reliability estimates the degree to which items (measured variables or relations) on remeasurement would order individuals responding to them in the same way.

Reliability measures can be devided into two major classes:

- *measures of stability* (e.g., test-retest, alternative form, true score measurement model)
- *measures of equivalence* (e.g., split-half coefficient, Cronbach's alpha, theta coefficient)

Estimating validity

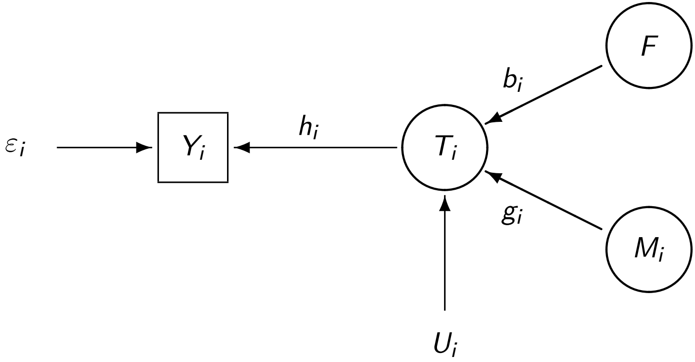
By validity we estimate whether or not one's items measure what they are intended to measure.

There are several approaches to estimate the validity:

- Criterion–related validity
- Content validity
- Construct validity (convergent validity and discriminant validity)
- Validity of a known group,
- Validity as non-method effect (true score measurement model)

True score measurement model

By the true score approach the reliability and the validity (Saris and Andrews 1991) of a single survey question (variable or relation) obtained by a specific method is estimated.



Reliability and validity

In the true score model, reliability is defined as the proportion of the variance in Y_i remaining stable across repetitions of the same measure, or:

$$reliability = \frac{var(T_i)}{var(Y_i)} = h_i^2$$

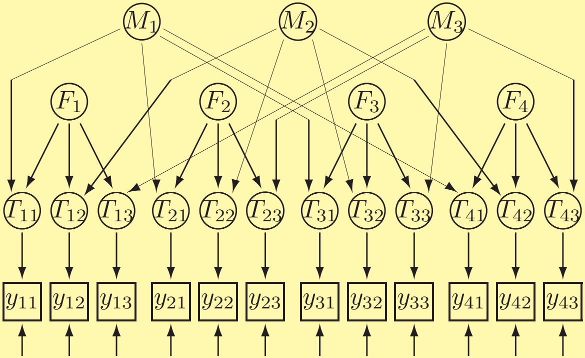
Validity is defined as the percentage of the variance of the stable component T_i explained by the variable of interest F, or:

$$validity = b_i^2$$

Estimation of reliability and validity

- In the true score model (with only one measurement) the coefficients h_i and b_i can not be estimated. It has been shown that at least three variables measured by at least three methods should be considered.
- In the next figure measurement model with four variables measured by three methods is presented.
- Using this model and structural equation modeling techniques, the reliability and the validity coefficients can be estimated for each measured variable.

An example of MTMM true score model



Reliabilty and validity of measuring whole networks

An experimental design to study systematically the impact of different measurement characteristics on the reliability and validity of whole network data in school classes (Ferligoj, Hlebec 1999; Hlebec, Ferligoj 2002) was the following:

- In the first phase of the study, estimates of reliability and validity were obtained for each relation in each of ten school classes, using the MTMM approach.
- In the second phase, the effects of the characteristics of the measurement instruments used in different classes were analyzed to explain the variability of the estimates for the reliability and validity. A meta analysis of MTMM results was done by multiple classification analysis (MCA).

Traits - measured relations

- exchange of study materials (*instrumental support*),
- exchange of information in the case of long-term illness (*informational support*),
- invitation to a birthday party (*social companionship*), and
- discussion of important personal matters (*emotional support*).

Relations were repeated in two ways:

- respondents described whom they would ask for a particular exchange (*original question*), and
- who would ask them for a particular exchange (*reversed question*).

Measurement scales and data collection technique

To measure the strength of ties, four measurement scales were used:

- a binary scale,
- a five-point ordinal scale,
- a five-point ordinal scale with labels, and
- a line drawing scale.

Two data collection technique were used:

- free recall,
- recognition (list),

Data collection

- There were ten classes: a class of university students, a class of pupils from a vocational school, and eight classes of pupils from a high school in Ljubljana.
- The first data collection (first class) was done in May 1993, the next one (second class) in May 1995 and the last one (the last eight classes) in January 1998.
- Paper-and-pencil data collection mode in all classes was used.
- Data were collected within one interview at intervals of approximately twenty minutes or after a week.
- In each class, only three scales were applied in keeping with traditional MTMM design. Within each class, the ordering of three selected scales, the time intervals between three repetitions, and the data collection method were varied.

Plan of the study

- First, the vectorization of all relational matrices for each class was performed.
- Then the reliability and the validity coefficients were estimated for all relations within each of the ten classes by MTMM approach.
- In the last phase, a meta-analysis was performed.

Some results

- Binary scale and the first presentation of measurement instruments give the least reliable measure.
- The most reliable measures were obtained by ordinal scales, among which the five-category ordinal scale with labels gave the most reliable measures.
- The two data collection methods (free recall and recognition) and the two types of network questions (original, reciprocated) yield equally reliable data.
- The time between repetitions is the most important predictor variable in the first meta-analysis: when a measure is presented first, it is the least reliable. When a measure is repeated after twenty minutes, its reliability estimate significantly increases.
- The measures of emotional and informational support are more reliable than those of material support and companionship.

Reliabilty and validity of measuring egocentered networks

- Studying the measurement quality of egocentered network measurement instruments is even more important, since data about the network and its characteristics and the characteristics of network members are usually given by the respondent (ego).
- The aim is to estimate the reliability and validity of frequently used name interpreters. As the unit of analysis is egocentered network as a whole, the variables are defined as averages of name interpreters for each egocentered network. Therefore, the reliability and validity of the averages for these variables were studied (Kogovšek et al. 2002; Kogovšek, Ferligoj 2004, 2005).
- Reliability and validity coefficients were estimated by the MTMM true score model.
- The effect of factors such as methods used and personal characteristics of respondents (egos) that can affect the quality of data was estimated by a meta analysis as before in the case of whole-networks.

By alters or by questions?

Name interpreter questions can be asked in two ways:

- *by alters* is to take each alter individually and to ask all questions about him/her, going alter by alter until the end of the list of alters;
- *by questions* is to take the question and ask this question to all alters on the list, going question by question until the end of the list.

Telehone or face-to-face mode?

It is expected that cognitively more demanding questions (e.g., frequency of contact between ego and his/her alters) are more prone to measurement errors in *telephone* than in *face-to-face or personal interviews*.
On the other hand, with the lack of the physical presence of the interviewer, telephone interviews may be more anonymous than personal interviews, which could produce more socially desirable responses to sensitive questions (e.g., feelings of closeness, frequency of alters upsetting the ego).

Split ballot MTMM design

The standard MTMM true score model requires respondent to answer the selected questions at least three times with different methods. This is a tedious task for respondents. Therefore, we decided to use a form of split ballot MTMM design (Saris, 1999) in which separate groups of respondents received different combinations of only two methods.

Study design

In the study three groups were used, each with two out of the three methods. The methods used were combinations of the data collection mode (telephone, face-to-face) and data collection approach (by alters, by questions):

- face-to-face/by alters
- telephone/by alters
- telephone/by questions

Group	N	First interview	Second interview
1	320	Face-to-face/by alters	Telephone/by alters
2	311	Face-to-face/by alters	Telephone/by questions
3	402	Telephone/by alters	Telephone/by questions

Data collection

- The data about five dimensions of social support and four properties of ties between ego and his/her alters were collected between March and June 2000 by computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI) for a representative sample of 1033 inhabitants of the city of Ljubljana.
- These respondents produced 7223 alters.
- The time span between the two measurements was one week.

Some results

- The data collection approach (by alters/by questions) mostly affected the reliability of measurement, whereas the data collection mode (telephone/face-to-face) mostly affected the validity of measurement.
- The telephone/by questions measurement method had a slightly higher validity than telephone/by alters, but had the worst reliability of all three methods.
- The personal interview by alters measurement method had relatively good reliability, but the worst validity.

Therefore, the telephone/by alters measurement method appears to be the optimal choice when measuring the characteristics of ties in egocentered networks.
The reason for this may lie in the relative sensitivity of the topic and the relative anonymity of the telephone method.

. . . Results

- Older respondents had lower reliability and lower validity of measures.
- Gender had a statistically significant effect only on the validity of measurement. Tie characteristics were, on average, more validly measured among males.
- The effects of education proved to be statistically non-significant.
- Consistent with personality theory, those who were both more extraverted and emotionally stable had a higher validity of measurement.

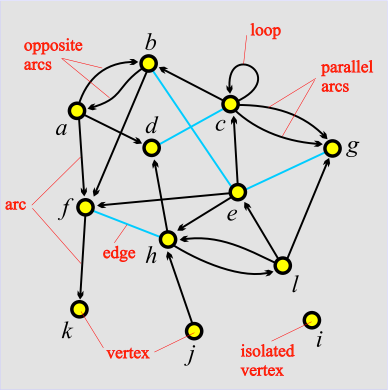
Measuring egocentered networks on the Web

- The layout of the web survey matters. E.g., different numbers of boxes to name the alters gave radically different network sizes (Vehovar et al. 2008).
- 'By questions' performed better than 'by alters' on reliability and validity of network measures using the MTMM approach on PhD students of three countries (Coromina and Coenders 2006). This is oposite result as in the case of telephone and face-to-face mode.
- Kogovšek (2006) compared web and telephone modes on a sample of students: the telephone mode produced more reliable data than the web mode. There were no large differences in the validity of measurement.

Some open problems

- To perform similar systematic studies on reliabilty and validity of whole-network measures on different populations (not only on students),
- To compare different types of modes (e.g., telephone, face-to-face, and web mode) in the same study.

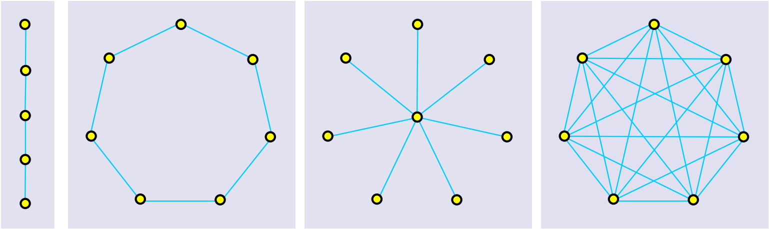
Graph theory concepts



A *graph* is based on two sets – set of *vertices* (nodes), that represent the selected *units*, and set of *lines* (links), that represent *ties* between units. A line can be *directed* – an *arc*, or *undirected* – an *edge*.

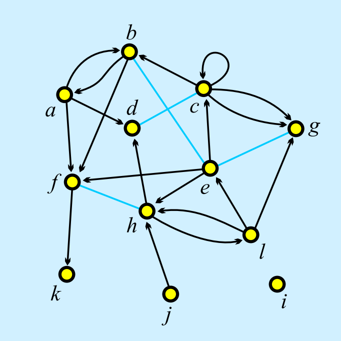
arc = directed line, (a, d)
a is the *initial* vertex,
d is the *terminal* vertex.
edge = undirected line, $(c : d)$
c and d are *end* vertices.

Special graphs – path, cycle, star, complete



Graphs: *path* P_5, *cycle* C_7, *star* S_8 in *complete graph* K_7.

Degrees



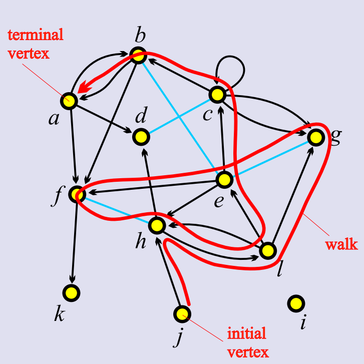
degree of vertex v, $\deg(v)$ = number of lines with v as end-vertex;

indegree of vertex v, $\mathrm{indeg}(v)$ = number of lines with v as terminal vertex (end-vertex of an edge is both initial and terminal);

outdegree of vertex v, $\mathrm{outdeg}(v)$ = number of lines with v as initial vertex.

$$n = 12,\ m = 23,\ \mathrm{indeg}(e) = 3,\ \mathrm{outdeg}(e) = 5,\ \deg(e) = 6$$

Walks



An example of a walk:

$s = (j, h, l, g, e, f, h, l, e, c, b, a)$

length $|s|$ of the walk s is the number of lines it contains.

$|s| = 11$

A walk is *closed* iff its initial and terminal vertex coincide.

If we don't consider the direction of the lines in the walk we get a *semiwalk* or *chain*.

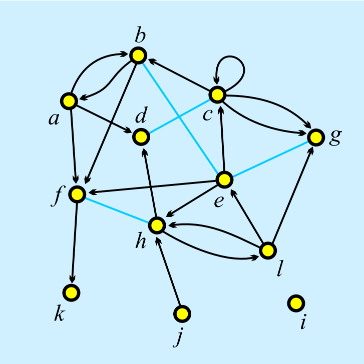
path – walk with all vertices different

trail – walk with all lines different

cycle – closed walk with all internal vertices different

A graph is *acyclic* if it doesn't contain any cycle.

Shortest paths



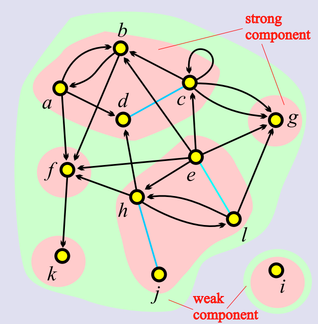
A shortest path from u to v is also called a *geodesic* from u to v. Its length is denoted by $d(u, v)$.

If there is no walk from u to v then $d(u, v) = \infty$.

$d(j, a) = |(j, h, d, c, b, a)| = 5$

$d(a, j) = \infty$

Connectivity



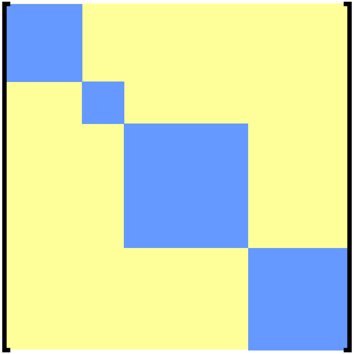
Vertex u is *reachable* from vertex v iff there exists a walk with initial vertex v and terminal vertex u.

A subset of the set of vertices is called a *strongly connected component* if (taking directions of lines into account) from each vertex of the subset we can reach every other vertex belonging to the same subset.

If direction of lines is not important, such a subset is called a *weakly connected component*.

Weak and strong connectivity are equivalence relations.

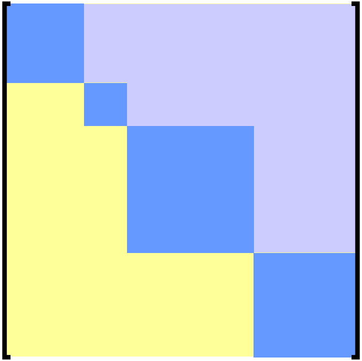
Weak components

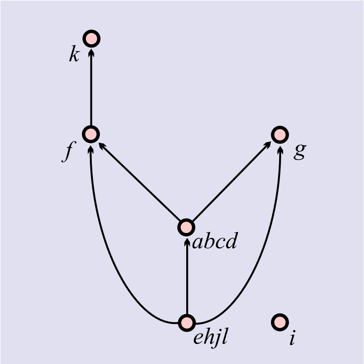


Reordering the vertices of network such that the vertices from the same class of weak partition are put together we get a matrix representation consisting of diagonal blocks – weak components.

Most problems can be solved separately on each component and afterward these solutions combined into final solution.

Reduction (condensation)





If we shrink every strong component of a given graph into a vertex, delete all loops and identify parallel arcs the obtained *reduced* graph is acyclic.

Important units in network

SNA

Ferligoj

The most important distinction between different *important unit measures* is based on whether the considered network is *directed or undirected*. This gives us two main types of measures:

- **undirected case**: measures of *centrality*.
  Example: A city is central if a lot of roads are passing through it.
- **directed case** with two subgroups of *prestige*:
  - measures of *support*, based on incoming arcs; and
  - measures of *influence*, based on out-going arcs.

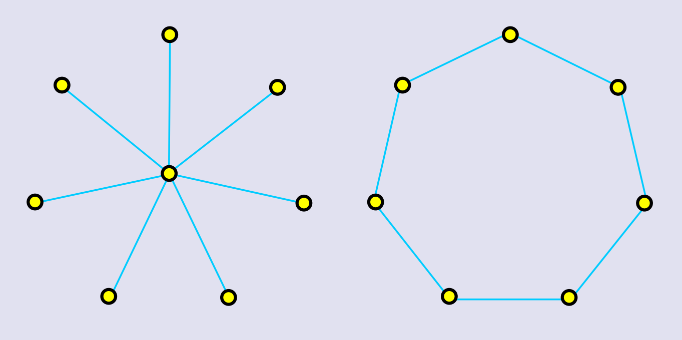
  Examples: A unit has high influence if (s)he gives commands to several others. A unit has high support if a lot of people vote for her/him.

Unit centrality measures

A unit is central if

- it has high *degree*,
- it is easy accessible or *close* to all other units,
- it lies on several geodesics (shortest paths) *between* units.

Star and cycle



According to all criteria mentioned, the central unit in the star is the most central, while all units in the cycle are equally central.

Degree measures

The simplest measure of importance is the degree of unit x

degree centrality	$c_D(x) = \deg(x)$
degree support	$c_{inD}(x) = \text{indeg}(x)$
degree influence	$c_{outD}(x) = \text{outdeg}(x)$

Such measures are called *absolute measures*, which cannot be compared accross different networks with different numbers of units.

Relative measures

Because of the comparabilty an absolute measure c has to be *normalized* in such a way that it is defined in the interval $[0, 1]$:

$$C(x) = \frac{c(x) - c_{min}}{c_{max} - c_{min}}$$

where c_{min} is the minimum value of the measure c and c_{max} is the maximal value of it. Such measures are called *relative measures*.

Relative degree measures

Relative degree centrality measure is:

$$C_D(x) = \frac{c_D(x)}{\text{highest degree}} = \frac{c_D(x)}{n-1}$$

and similary *relative degree support and influence measures* are:

$$C_{inD}(x) = \frac{c_{inD}(x)}{n}$$

$$C_{outD}(x) = \frac{c_{outD}(x)}{n}$$

Closeness measures

Sabidussi (1966) introduced a measure of centrality according to the closeness of unit x to all others:

$$c_C(x) = \frac{1}{\sum_{y \in \mathcal{U}} d(x, y)}$$

where $d(x, y)$ is the distance of shortest path between unit x and y, and $\mathcal{U}$ is the set of all units.
The most central units, according to the closeness centrality measure, can quickly interact with all others because they are close to all others.

. . . Closeness measures

Closeness can be measured also for the directed networks. In this case the two measures are:

- *closeness measure of influence* $c_{inC}(x)$ and
- *closeness measure of support* $c_{outC}(x)$

Relative closeness measures

Sabidussi (1966) introduced a *relative closeness measure*:

$$C_C(x) = \frac{n-1}{\sum_{y \in \mathcal{U}} d(x,y)}$$

Closeness measure can be computed for undirected and directed networks.
There are two possibilities for directed networks:

- incoming ties: in how many steps the selected unit can be reached from all others (*closeness support measure*);
- outgoing ties: in how many steps all other units can be reached from the selected one (*closeness influence measure*).

Betweeness measures

The units that can control the information flow in the network are also important.
If we assume that the information flow uses only the shortest paths (geodesics) we get a measure of *betweeness* (Anthonisse 1971, Freeman 1977):

$$c_B(x) = \sum_{y<z} \frac{\#\text{ of shortest paths between units } y \text{ and } z \text{ through unit } x}{\#\text{ of shortest path between units } y \text{ and } z}$$

For directed networks similar betweeness measure can be defined.

Relative betweeness measures

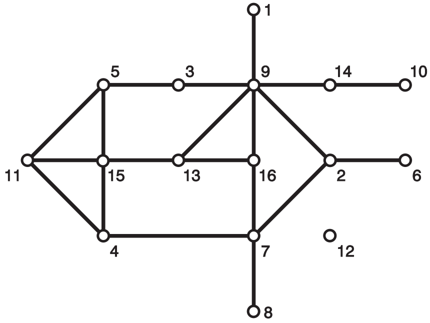
Relative betweeness centrality measure is:

$$C_B(x) = \frac{c_B(x)}{(n-1)(n-2)/2}$$

For directed networks the relative betweeness measure is:

$$C_B(x) = \frac{c_B(x)}{(n-1)(n-2)}$$

Example 1: Marriage ties of Florentine families



1. Acciaiuoli
2. Albizzi
3. Barbadori
4. Bischeri
5. Castellani
6. Ginori
7. Guadagni
8. Lamberteschi
9. Medici
10. Pazzi
11. Peruzzi
12. Pucci
13. Ridolfi
14. Salviati
15. Strozzi
16. Tornabuoni

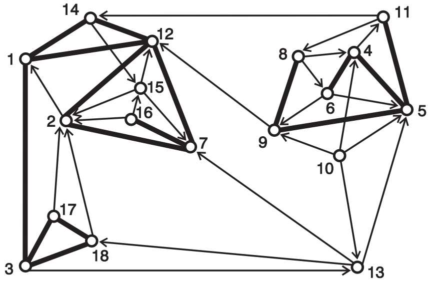
SNA

Ferligoj

Centrality measures

	C_D	C_C	C_B
1. Acciaiuoli	0.071	0.368	0.000
2. Albizzi	0.214	0.483	0.212
3. Barbadori	0.143	0.438	0.093
4. Bischeri	0.214	0.400	0.104
5. Castellani	0.214	0.389	0.055
6. Ginori	0.071	0.333	0.000
7. Guadagni	0.286	0.467	0.255
8. Lamberteschi	0.071	0.326	0.000
9. Medici	**0.429**	**0.560**	**0.522**
10. Pazzi	0.071	0.286	0.000
11. Peruzz i	0.214	0.368	0.022
12. Ridolfi	0.214	0.500	0.114
13. Salviati	0.143	0.389	0.143
14. Strozzi	0.286	0.438	0.103
15. Tornabuoni	0.214	0.483	0.092

Example 2: Sampson monastery – liking



1. John Bosco
2. Gregory
3. Basil
4. Peter
5. Bonaventura
6. Berthold
7. Mark
8. Victor
9. Ambrose
10. Romuald
11. Louis
12. Winfrid
13. Amand
14. Hugh
15. Boniface
16. Albert
17. Elias
18. Simplicius

Prestige measures

	C_{inD}	C_{inC}	C_B
1. John Bosco	0.235	0.472	0.332
2. Gregory	**0.353**	0.486	0.122
3. Basil	0.176	0.378	**0.358**
4. Peter	0.235	0.293	0.065
5. Bonaventura	**0.353**	0.370	0.227
6. Berthold	0.118	0.239	0.007
7. Mark	0.294	0.486	0.079
8. Victor	0.118	0.239	0.013
9. Ambrose	0.235	0.293	0.146
10. Romuald	0.000	0.000	0.000
11. Louis	0.118	0.283	0.089
12. Winfrid	**0.353**	**0.586**	0.220
13. Amand	0.118	0.298	0.250
14. Hugh	0.177	0.447	0.111
15. Boniface	0.118	0.354	0.016
16. Albert	0.059	0.333	0.011
17. Elias	0.118	0.293	0.002
18. Simplicius	0.176	0.304	0.021

Hubs and authorities

In the directed networks another two types of important units can be identified: hubs and authorities.

To each unit x of a network we assign two values: quality of its content (*authority*) u_x and quality of its references (*hub*) v_x.

The unit x is a better authority than unit y, if $u_x > u_y$.

A good authority is selected by good hubs; and good hub points to good authorities (see Kleinberg, 1998):

$$u_x = \sum_y v_y \qquad \text{and} \qquad v_y = \sum_x u_x$$

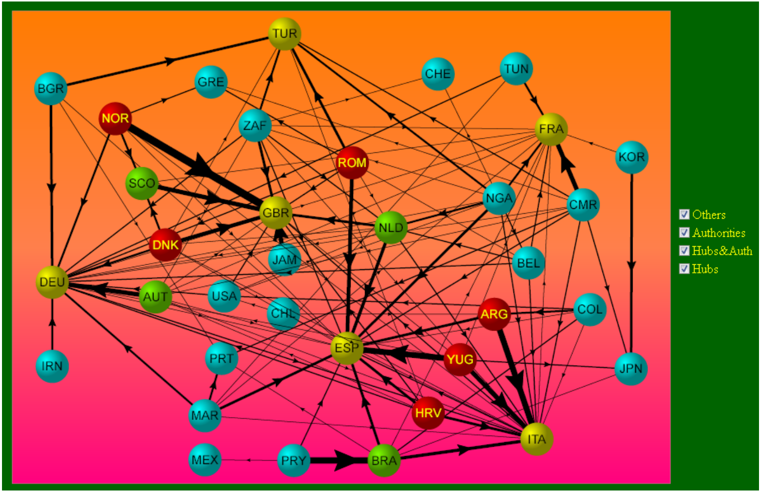
The best authority has the largest weight u and the best hub has the largest weight v. They are calculated iteratively.

Example

The network consists of the 32 football natonal teams which participated in the World Championship in Paris, 1998. Players of the national teams often have contracts in the other countries. This constitutes a players market where national teams export players to the other countries. The number of 'exported' players of a national team to a country defines a value on this tie. 32 national teams had contracts in 35 countries. The network is highly asymmetric: some countries only export players, some countries are only importers.

The best authorities (importers - countries that import good players) are denoted by yellow, the best hubs (exporters - countries that export players in the best leags) are denoted by red.

World Championship in Paris, 1998



Other measures

- For large networks fast algorithms for computing centrality and prestige measures were proposed (e.g., Brandes, 2001; Brandes and Pich, 2007; Becchetti st al., 2008; Guha and McGregor, 2012; Lambertini et al., 2014).
- Eigenvector centrality measures were defined also for valued networks (e.g., Katz, 1953; Bonacich, 1991).
- Generalizations of centrality and prestige measures to valued networks were proposed (e.g., Opsahl et al., 2010).
- Specific centrality and prestige measures were defined in different scientific fields.

Clusters, clusterings, partitions, hierarchies

A nonempty subset $C \subseteq \mathcal{U}$ is called a *cluster* (group). A nonempty set of clusters $\mathbf{C} = \{C_i\}$ forms a *clustering*. Clustering $\mathbf{C} = \{C_i\}$ is a *partition* iff

$$\cup\mathbf{C} = \bigcup_i C_i = \mathcal{U} \quad \text{and} \quad i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

Clustering $\mathbf{C} = \{C_i\}$ is a *hierarchy* iff

$$C_i \cap C_j \in \{\emptyset, C_i, C_j\}$$

Hierarchy $\mathbf{C} = \{C_i\}$ is *complete*, iff $\cup\mathbf{C} = \mathcal{U}$; and is *basic* if for all $x \in \cup\mathbf{C}$ also $\{x\} \in \mathbf{C}$.

Contraction of cluster

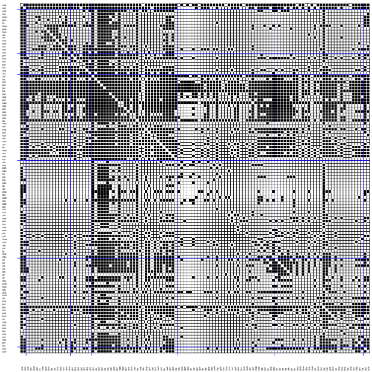
Contraction of cluster C is called a graph $\mathcal{G}/C$, in which all vertices of the cluster C are replaced by a single vertex, say c. In a network over graph $\mathcal{G}$ we have also to specify how are determined the values/weights in the shrunk part of the network. Usually as the sum or maksimum/minimum of the original values.

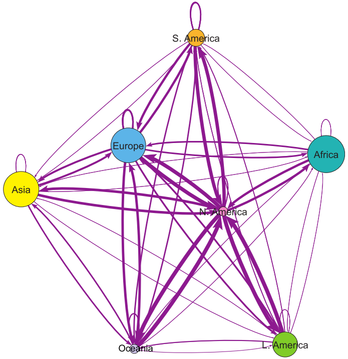
Example: Snyder and Kick World Trade

The network consists of trade relations (118 countries - vertices, 515 arcs, 2116 edges). The source of the data is the paper of Snyder, David and Edward Kick (1979): *The World System and World Trade: An Empirical Exploration of Conceptual Conflicts*, Sociological Quaterly, 20,1, 23-36.

The data about the (sub)continents partition is also given: 1 - Europe, 2 - North America, 3 - Latin America, 4 - South America, 5 - Asia, 6 - Africa, 7 - Oceania.

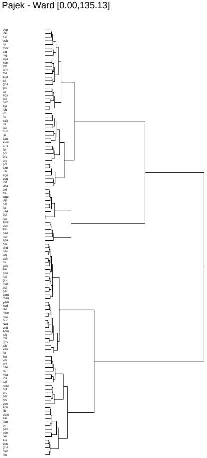
Contracted clusters – world trade

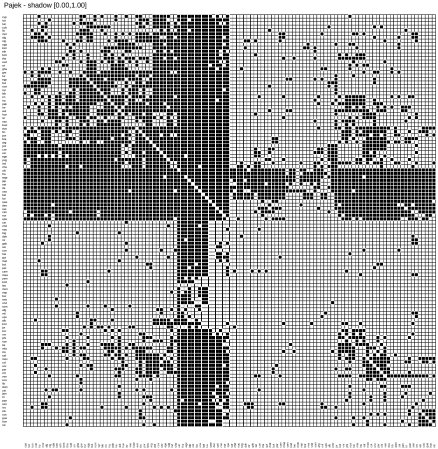




Snyder and Kick's world trade. Matrix display of dense networks.

Clustering

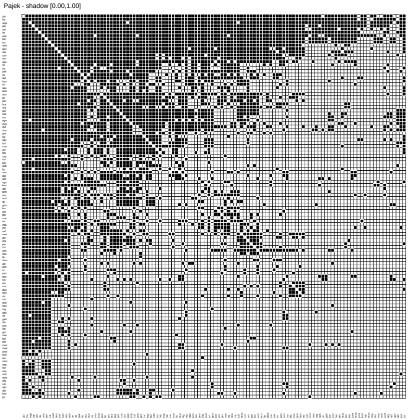




Reordering clustering

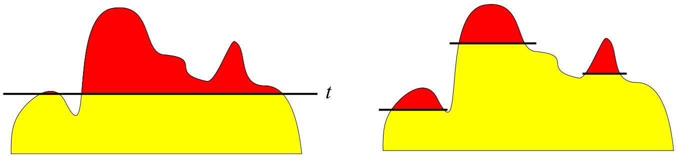
The order of clusters in a hierarchy is not fixed and can be changed.

We see the typical core–periphery structure.



Islands

If we represent a given or computed property of vertices / lines as a height of vertices / lines and we immerse the network into a water up to selected property level we get *cuts*. Varying the level we get different *islands* (connected subnetworks).



Batagelj et al. (2003) developed very efficient algorithms to determine the islands hierarchy and to list all the islands of selected sizes. An island is *simple* iff it has a single peak.

. . . Islands

Islands are very general and efficient approach to determine the 'important' subnetworks in a given network.
We have to express the goals of our analysis with a related property of the vertices or weight of the lines. Using this property we determine the islands of an appropriate size (in the interval k to K).
In large networks we can get many islands which we have to inspect individually and interpret their content.
An important property of islands is that they identify locally important subnetworks on different levels.

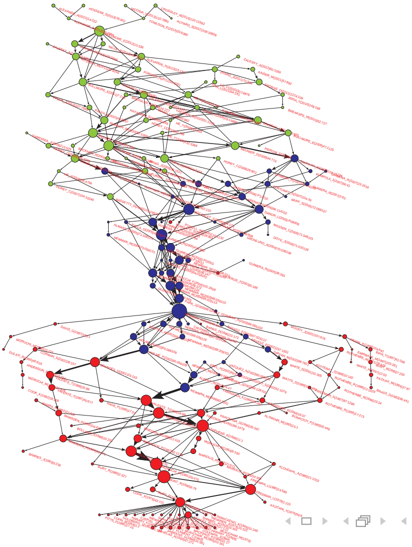
Example: The centrality citation network

Citation network of all papers dealing with the centrality literature till 2013 (WoS) was analyzed by Batagelj, Doreian, Ferligoj, Kejžar (2014). There are 995,783 scientific papers with 1,856,102 citation links. This is a binary almost acyclic temporal network.

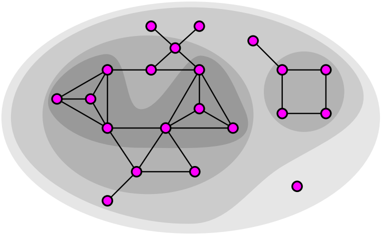
Weights on arcs have to be defined to find the most important subnetworks. Hummon and Doreian (1990) proposed a weight that is proportional to the number of all paths passing through the arc (Search Path Count - SPC). For these weights the main line-island was obtained.

The main line-island

red, SNA; blue, physics; green, neuroscience



Cores



The notion of core was introduced by Seidman in 1983. A subset of the set of vertices is called a k-core if every vertex from the subset is connected to at least k vertices from the same set.

The core of maximum order is also called the *main* core.

The *core number* of vertex v is the highest order of a core that contains this vertex.

By cores we can efficiently determine dense groups in the network.

Properties of cores

From the figure, representing 0, 1, 2 and 3 core, we can see the following properties of cores:

- The cores are nested
- Cores are not necessarily connected subgraphs.

The cores, because they can be determined very efficiently, are one among few concepts that provide us with meaningful decompositions of large networks.

Generalized cores

Cores can be defined also for valued networks. Here, we define a *p-function* on the vertex which is defined on the values of the lines going to the vertex neighbors.

p_S-finction is defined by sum:

$$p_S(v) = \sum_{u \in N(v)} w(v, u)$$

where $N(v)$ are neighbors of the vertex v and $w(v, u)$ value on the line between verteces v and u.

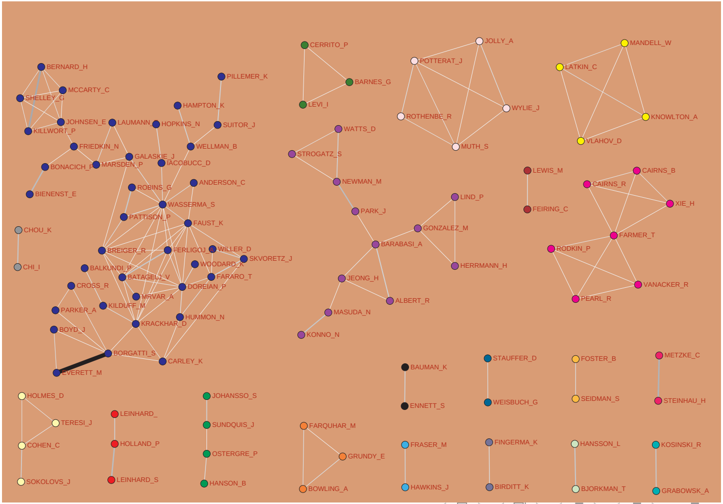
The *p_S-core* at the level t means that each vertex in the core has the value of the p_S function at least t inside the core and the obtained cluster is a maximal one.

Example: SNA co-authorship network

Batagelj and collegues (2014) studied the co-authorship network of the SNA community in 2008. It is a valued network where values were normalized by the number of the co-authors. On this network p_S-core at the level 0.75 was computed.

The largest cores component consists of the main stream social networks researchers with the most intensive collaboration pairs. The second largest component consists of physicists with more intensive collaboration pairs.

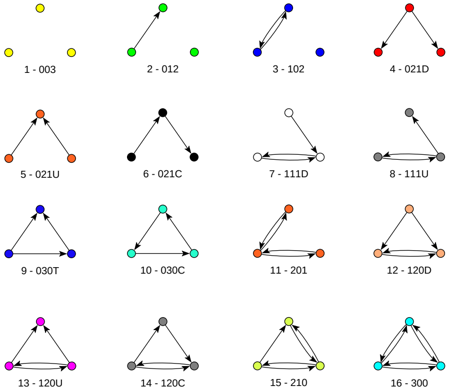
p_S-cores in SNA co-authorship network



Pattern searching

If a selected *pattern* determined by a given graph does not occur frequently in a sparse network the straightforward backtracking algorithm applied for pattern searching finds all appearences of the pattern very fast even in the case of very large networks. Pattern searching was successfully applied to searching for patterns of atoms in molecula (carbon rings) and searching for relinking marriages in genealogies.

Triads



Let $\mathcal{G} = (\mathcal{V}, R)$ be a simple directed graph without loops.
A *triad* is a subgraph induced by a given set of three vertices.
There are 16 types of triads.

Triadic spectrum

Triad:	BA	CL	RC	R2C	TR	HC	39+	p1	p2	p3	p4
003		✓	✓		✓	✓				✓	✓
012					✓	✓	✓			✓	✓
102	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
021D			✓	✓	✓	✓	✓				✓
021U			✓	✓	✓	✓	✓			✓	✓
021C									✓		✓
111D											✓
111U								✓	✓		
030T			✓	✓	✓	✓	✓		✓		✓
030C								✓	✓		✓
201											
120D			✓	✓	✓	✓	✓			✓	✓
120U			✓	✓	✓	✓	✓	✓	✓		✓
120C							✓		✓		
210						✓	✓		✓		
300	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Triad Micro-Models:
BA: Ballance (Cartwright and Harary, '56)
CL: Clustering Model (Davis. '67)
RC: Ranked Cluster (Davis & Leinhardt, '72)
R2C: Ranked 2-Clusters (Johnsen, '85)
TR: Transitivity (Davis and Leinhardt, '71)
HC: Hierarchical Cliques (Johnsen, '85)
39+: Model that fits D&L's 742 mats N :39-72
p1-p4: Johnsen, 1986. Process Agreement Models.

Moody

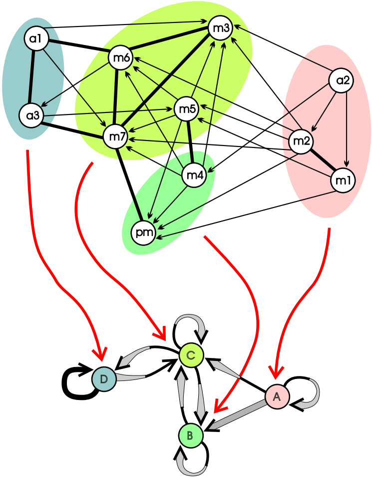
Several properties of a graph can be expressed in terms of its *triadic spectrum* – distribution of all its triads. It also provides ingredients for p^* network models.

A direct approach to determine the triadic spectrum is of order $O(n^3)$; but in most large graphs it can be determined much faster.

Blockmodeling

The goal of *blockmodeling* is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily.

Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, according to some *meaningful* definition of equivalence.



Cluster, clustering, blocks

One of the main procedural goals of blockmodeling is to identify, in a given network $\mathbf{N} = (\mathcal{U}, R)$, $R \subseteq \mathcal{U} \times \mathcal{U}$, *clusters* (classes) of units that share structural characteristics defined in terms of R. The units within a cluster have the same or similar connection patterns to other units. They form a *clustering* $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ which can be a *partition* of the set $\mathcal{U}$. Each partition determines an equivalence relation (and vice versa). Let us denote by $\sim$ the relation determined by partition $\mathbf{C}$.

A clustering $\mathbf{C}$ partitions also the relation R into *blocks*

$$R(C_i, C_j) = R \cap C_i \times C_j$$

Each such block consists of units belonging to clusters C_i and C_j and all arcs leading from cluster C_i to cluster C_j. If $i = j$, a block $R(C_i, C_i)$ is called a *diagonal* block.

The Everett network

	a	b	c	d	e	f	g	h	i	j
a	0	1	1	1	0	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0	0
c	1	1	0	1	0	0	0	0	0	0
d	1	0	1	0	1	0	0	0	0	0
e	0	1	0	1	0	1	0	0	0	0
f	0	0	0	0	1	0	1	0	1	0
g	0	0	0	0	0	1	0	1	0	1
h	0	0	0	0	0	0	1	0	1	1
i	0	0	0	0	0	1	0	1	0	1
j	0	0	0	0	0	0	1	1	1	0

	a	c	h	j	b	d	g	i	e	f
a	0	1	0	0	1	1	0	0	0	0
c	1	0	0	0	1	1	0	0	0	0
h	0	0	0	1	0	0	1	1	0	0
j	0	0	1	0	0	0	1	1	0	0
b	1	1	0	0	0	0	0	0	1	0
d	1	1	0	0	0	0	0	0	1	0
g	0	0	1	1	0	0	0	0	0	1
i	0	0	1	1	0	0	0	0	0	1
e	0	0	0	0	1	1	0	0	0	1
f	0	0	0	0	0	0	1	1	1	0

	A	B	C
A	1	1	0
B	1	0	1
C	0	1	1

Equivalences

Regardless of the definition of equivalence used, there are two basic approaches to the equivalence of units in a given network (Faust, 1988):

- the equivalent units have the same connection pattern to the **same** neighbors;
- the equivalent units have the same or similar connection pattern to (possibly) **different** neighbors.

The first type of equivalence is formalized by the notion of structural equivalence and the second by the notion of regular equivalence with the latter a generalization of the former.

Structural equivalence

Units are equivalent if they are connected to the rest of the network in *identical* ways (Lorrain and White, 1971). Such units are said to be *structurally equivalent*.

In other words, x and y are structurally equivalent iff:

s1. $\mathrm{x} R \mathrm{y} \Leftrightarrow \mathrm{y} R \mathrm{x}$

s2. $\mathrm{x} R \mathrm{x} \Leftrightarrow \mathrm{y} R \mathrm{y}$

s3. $\forall \mathrm{z} \in \mathcal{U} \setminus \{\mathrm{x}, \mathrm{y}\} : (\mathrm{x} R \mathrm{z} \Leftrightarrow \mathrm{y} R \mathrm{z})$

s4. $\forall \mathrm{z} \in \mathcal{U} \setminus \{\mathrm{x}, \mathrm{y}\} : (\mathrm{z} R \mathrm{x} \Leftrightarrow \mathrm{z} R \mathrm{y})$

. . . Structural equivalence

The blocks for structural equivalence are null or complete with variations on diagonal in diagonal blocks.

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}$$

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{matrix} \qquad \begin{matrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{matrix}$$

Regular equivalence

Integral to all attempts to generalize structural equivalence is the idea that units are equivalent if they link in equivalent ways to other units that are also equivalent.

White and Reitz (1983): The equivalence relation $\approx$ on $\mathcal{U}$ is a *regular equivalence* on network $\mathbf{N} = (\mathcal{U}, R)$ if and only if for all $\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{w} \in \mathcal{U}$, $\mathrm{x} \approx \mathrm{y}$ implies both

R1. $\mathrm{x}R\mathrm{z} \Rightarrow \exists \mathrm{w} \in \mathcal{U} : (\mathrm{y}R\mathrm{w} \wedge \mathrm{w} \approx \mathrm{z})$

R2. $\mathrm{z}R\mathrm{x} \Rightarrow \exists \mathrm{w} \in \mathcal{U} : (\mathrm{w}R\mathrm{y} \wedge \mathrm{w} \approx \mathrm{z})$

. . . Regular equivalence

The blocks for regular equivalence are null or 1-covered blocks (Batagelj, Doreian, Ferligoj, 1992):

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{matrix}$$

Establishing blockmodels

The problem of establishing a partition of units in a network in terms of a selected type of equivalence is a special case of *clustering problem* that can be formulated as an optimization problem (Φ, P) as follows:

Determine the clustering $\mathbf{C}^\star \in \Phi$ *for which*

$$P(\mathbf{C}^\star) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where Φ is the set of *feasible clusterings* and P is a *criterion function*.

Criterion function

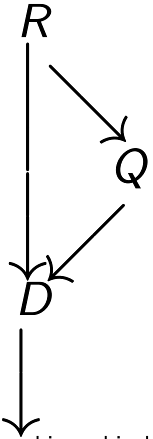
SNA

Ferligoj

Criterion functions can be constructed

- *indirectly* as a function of a *compatible* (dis)similarity measure between pairs of units, or
- *directly* as a function measuring the *fit* of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the considered types of connections (equivalence).

Indirect approach



RELATION

DESCRIPTIONS OF UNITS

original relation
path matrix
triads

DISSIMILARITY MATRIX

STANDARD CLUSTERING ALGORITHMS

hierarchical algorithms,
relocation algorithm, k-means algorithm, etc.

Dissimilarities

The dissimilarity measure d is *compatible* with a considered equivalence $\sim$ if for each pair of units holds

$$x_i \sim x_j \Leftrightarrow d(x_i, x_j) = 0$$

Not all dissimilarity measures typically used are compatible with structural equivalence. For example, the *corrected Euclidean-like dissimilarity*

$$d(x_i, x_j) = \sqrt{(r_{ii} - r_{jj})^2 + (r_{ij} - r_{ji})^2 + \sum_{\substack{s=1 \\ s \neq i,j}}^{n} ((r_{is} - r_{js})^2 + (r_{si} - r_{sj})^2)}$$

is compatible with structural equivalence.

The indirect clustering approach does not seem suitable for establishing clusterings in terms of regular equivalence since there is no evident way how to construct a compatible (dis)similarity measure.

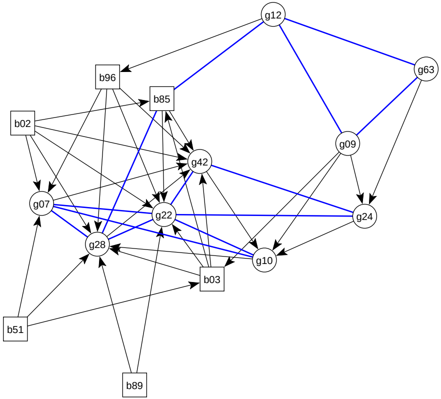
Example: Support network among informatics students

The analyzed network consists of social support exchange relation among fifteen students of the Social Science Informatics fourth year class (2002/2003) at the Faculty of Social Sciences, University of Ljubljana. Interviews were conducted in October 2002.

Support relation among students was identified by the following question:

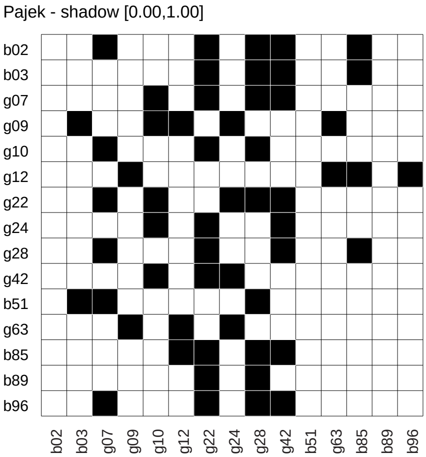
> *Introduction: You have done several exams since you are in the second class now. Students usually borrow studying material from their colleagues.*
> *Enumerate (list) the names of your colleagues that you have most often borrowed studying material from. (The number of listed persons is not limited.)*

Class network - graph

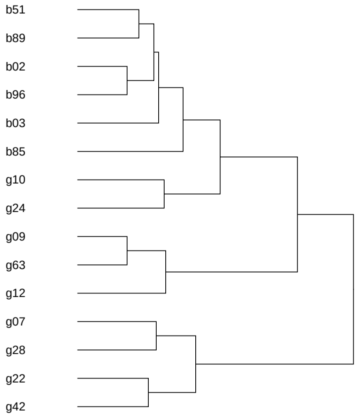


Vertices represent students in the class: circles – girls, squares – boys.
Reciprocated arcs are represented by edges.

Class network – matrix



Indirect approach



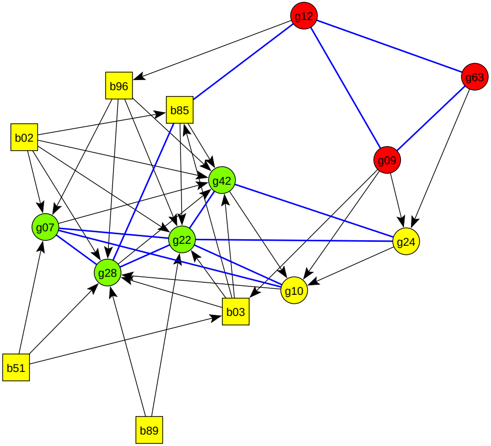
Using *Corrected Euclidean-like dissimilarity* and *Ward clustering method* we obtain the following dendrogram.

From it we can determine the number of clusters: 'Natural' clusterings correspond to clear 'jumps' in the dendrogram.

If we select 3 clusters we get the partition **C**.

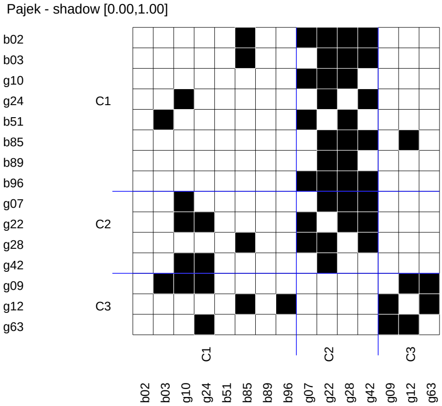
$$\mathbf{C} = \{\{b51, b89, b02, b96, b03, b85, g10, g24\}, \{g09, g63, g12\}, \{g07, g28, g22, g42\}\}$$

Partition into three clusters (Indirect approach)



On the picture, vertices in the same cluster are of the same color.

Matrix



The partition can be used also to reorder rows and columns of the matrix representing the network. Clusters are divided using blue vertical and horizontal lines.

Direct approach

The second possibility for solving the blockmodeling problem is to construct an appropriate criterion function directly and then use a local optimization algorithm to obtain a 'good' clustering solution.

Criterion function $P(\mathbf{C})$ has to be *sensitive* to considered equivalence:

$$P(\mathbf{C}) = 0 \Leftrightarrow \mathbf{C} \text{ defines considered equivalence.}$$

Criterion function

One of the possible ways of constructing a criterion function that directly reflects the considered equivalence is to measure the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the considered equivalence.

Given a clustering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, let $\mathcal{B}(C_u, C_v)$ denote the set of all ideal blocks corresponding to block $R(C_u, C_v)$.

Then the global error of clustering $\mathbf{C}$ can be expressed as

$$P(\mathbf{C}) = \sum_{C_u, C_v \in \mathbf{C}} \min_{B \in \mathcal{B}(C_u, C_v)} d(R(C_u, C_v), B)$$

where the term $d(R(C_u, C_v), B)$ measures the difference (error) between the block $R(C_u, C_v)$ and the ideal block B. d is constructed on the basis of characterizations of types of blocks. The function d has to be compatible with the selected type of equivalence.

Example

Empirical blocks | Ideal blocks

	a	b	c	d	e	f	g		a	b	c	d	e	f	g
a	0	1	1	0	1	0	0	a	0	1	1	0	0	0	0
b	1	0	1	0	0	0	0	b	1	0	1	0	0	0	0
c	1	1	0	0	0	0	0	c	1	1	0	0	0	0	0
d	1	1	1	0	0	0	0	d	1	1	1	0	0	0	0
e	1	1	1	0	0	0	0	e	1	1	1	0	0	0	0
f	1	1	1	0	1	0	1	f	1	1	1	0	0	0	0
g	0	1	1	0	0	0	0	g	1	1	1	0	0	0	0

Number of inconsistencies for each block

	A	B
A	0	1
B	1	2

$P = 4.$

Local optimization

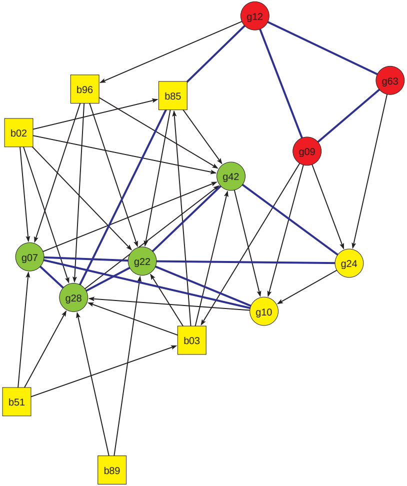
For solving the blockmodeling problem the relocation algorithm can be used:

Determine the initial clustering $\mathcal{C}$;
repeat:
  if in the neighborhood of the current clustering $\mathcal{C}$
  there exists a clustering $\mathcal{C}'$ such that $P(\mathcal{C}') < P(\mathcal{C})$
  then move to clustering $\mathcal{C}'$.

The neighborhood in this local optimization procedure is determined by the following two transformations:

- *moving* a unit x_k from cluster C_p to cluster C_q (*transition*);
- *interchanging* units x_u and x_v from different clusters C_p and C_q (*transposition*).

Partition into three clusters: Direct solution

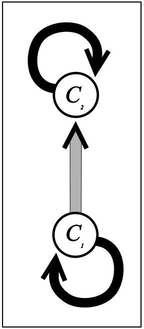


This is the same partition and has the number of inconsistencies.

Generalized blockmodeling

1	1	1	1	1	1	0	0
1	1	1	1	0	1	0	1
1	1	1	1	0	0	1	0
1	1	1	1	1	0	0	0
0	0	0	0	0	1	1	1
0	0	0	0	1	0	1	1
0	0	0	0	1	1	0	1
0	0	0	0	1	1	1	0

	C_1	C_2
C_1	complete	regular
C_2	null	complete



Generalized equivalence / block types

	Y				
	1	1	1	1	1
X	1	1	1	1	1
	1	1	1	1	1
	1	1	1	1	1

complete

	Y				
	0	1	0	0	0
X	1	1	1	1	1
	0	0	0	0	0
	0	0	0	1	0

row-dominant

	Y				
	0	0	1	0	0
X	0	0	1	1	0
	1	1	1	0	0
	0	0	1	0	1

col-dominant

	Y				
	0	1	0	0	0
X	1	0	1	1	0
	0	0	1	0	1
	1	1	0	0	0

regular

	Y				
	0	1	0	0	0
X	0	1	1	0	0
	1	0	1	0	0
	0	1	0	0	1

row-regular

	Y				
	0	1	0	1	0
X	1	0	1	0	0
	1	1	0	1	1
	0	0	0	0	0

col-regular

	Y				
	0	0	0	0	0
X	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0

null

	Y				
	0	0	0	1	0
X	0	0	1	0	0
	1	0	0	0	0
	0	0	0	1	0

row-functional

	Y			
	1	0	0	0
	0	1	0	0
X	0	0	1	0
	0	0	0	0
	0	0	0	1

col-functional

Pre-specified blockmodeling

In the previous slides the inductive approaches for establishing blockmodels for a set of social relations defined over a set of units were discussed. Some form of equivalence is specified and clusterings are sought that are consistent with a specified equivalence.

Another view of blockmodeling is deductive in the sense of starting with a blockmodel that is specified in terms of substance prior to an analysis.

Types of pre-specified blockmodels

The pre-specified blockmodeling starts with a blockmodel specified, in terms of substance, **prior to an analysis**. Given a network, a set of ideal blocks is selected, a family of reduced models is formulated, and partitions are established by minimizing the criterion function.

The basic types of blockmodels are:

*	*	*
*	0	0
*	0	0

core - periphery

*	0	0
*	*	0
?	*	*

hierarchy

*	0	0
0	*	0
0	0	*

clustering

Example 1: Class network

We expect that core-periphery model exists in the class network: some students having good studying material, some not.

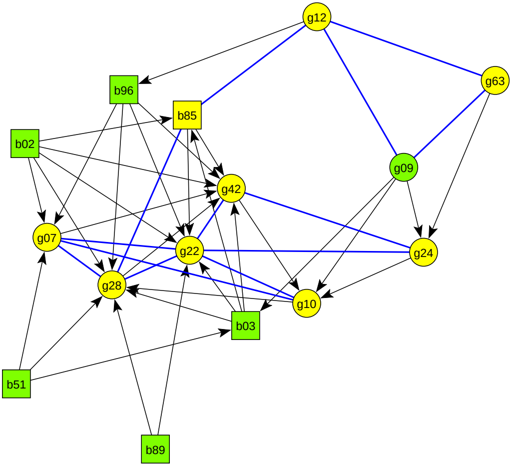
Pre-specified blockmodel: (com/complete, reg/regular, -/null block)

	1	2
1	[com reg]	-
2	[com reg]	-

Using local optimization we get the partition:

$$\mathbf{C} = \{\{b02, b03, b51, b85, b89, b96, g09\}, \{g07, g10, g12, g22, g24, g28, g42, g63\}\}$$

2 Clusters Solution



Model

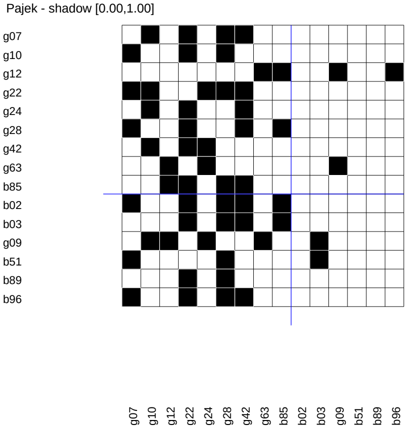


Image and error matrices:

	1	2
1	reg	-
2	reg	-

	1	2
1	0	3
2	0	2

Total error $= 5$

core-periphery

Example 2: Co-authorship networks of Slovenian researchers

Blockmodeling of vauled networks

	1	7	13	2	5	6	10	11	12	3	4	8	9
1									3		15	1	8
7										6	14	14	6
13					8						1	8	3
2		5					10	1	3	2	3	5	10
5			3			5		5			16	7	16
6			1		4			7	3	1		7	3
10				16		1		1	2	2	16	16	
11					2	2			2	2	8	5	14
12	2	2		2	2	2	2	11		8	2	2	6
3											19	3	1
4	2				1			1		6		1	19
8											5		6
9											19	1	

Žiberna (2007) proposed several approaches to generalized blockmodeling of valued networks, where values of the ties are assumed to be measured on at least interval scale. One of them is homogeneity blockmodeling. The basic idea of homogeneity blockmodeling is that the inconsistency of an empirical block with its ideal block can be measured by within block variability of appropriate values.

SNA

Ferligoj

Blockmodeling of three-way networks

The indirect approach to *structural equivalence blockmodeling in 3-mode networks* was proposed by Batagelj et al. (2007).

A *3-mode network* **N** over the basic sets X, Y and Z is determined by a ternary relation $R \subseteq X \times Y \times Z$. The notion of structural equivalence depends on which of the sets X, Y and Z are (considered) the same.

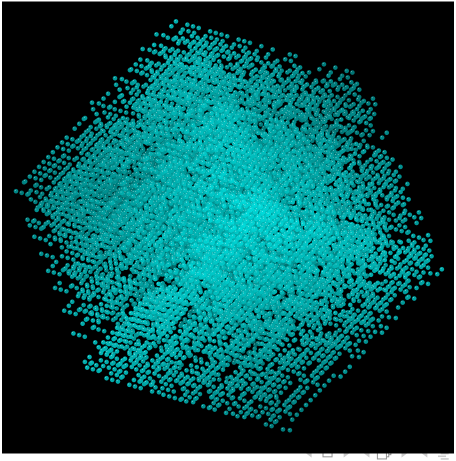
There are three basic cases:

- all three sets are different
- two sets are the same
- all three sets are the same

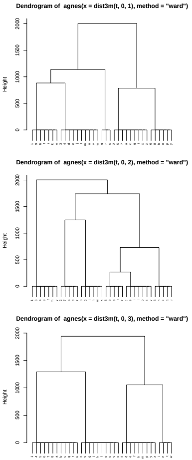
Example: Artificial dataset

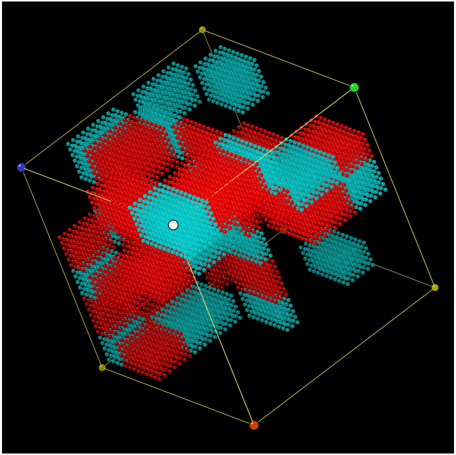
Randomly generated ideal structure

`rndTest(c(5,6,4),c(35,35,35))`:



Example: Solutions





Open problems in generalized blockmodeling

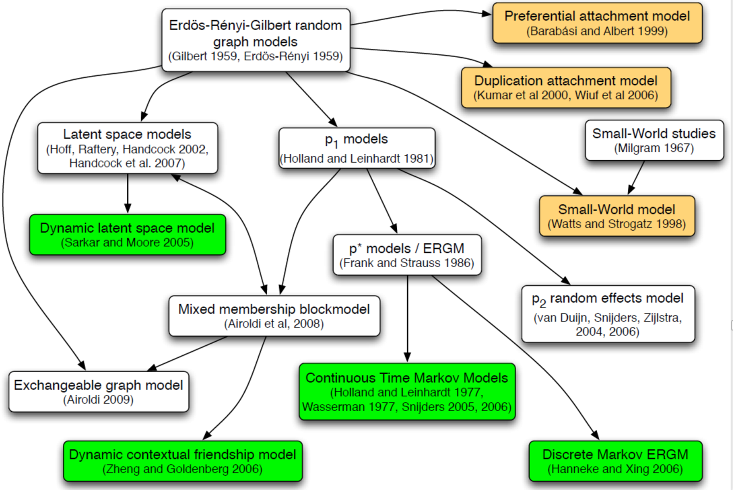
- Boundary problems
- Measurement errors
- Assesing fits of blockmodels
- Blockmodeling large networks
- Number of positions
- Dynamic blockmodels

See more in Ferligoj, Doreian, Batagelj (2011).

Statistical models

Goldenberg, Zheng, Fienberg, Airoldi, 2009

(white - static models, yellow - pseudo-dynamic, green - dynamic models)



Stochastic blockmodeling

Two units are stochastically equivalent if they have the same probability distribution of teir ties to other units (Holland et al., 1983; Wasserman, Andreson, 1987; Andreson et al., 1992).

- Nowicki and Snijders (2001) developed a Bayesian approach: units were diads and parameters were estimated by a Markov Chain Monte Carlo procedure.
- Airoldi et al. (2007) introduced a family of stochastic blockmodels that combine features of mixed-membership models and blockmodels in hierarchical Bayesian framework.
- Handcock et al. (2007) proposed a new model with latent positions under which the probability of a tie between two units depends on the distance between them in an unobserved Euclidean "social space". They proposed two estimation procedures: a two-stage maximum likelihood method and a fully Bayesian method.

Exponential Random Graph Models (ERGM) or p^* models

Formally a *random graph* Y consists of a set of n units and m dyads (ties) $\{Y_{ij} : i = 1, \ldots, n; j = 1, \ldots, n\}$ where $Y_{ij} = 1$ if the units (i, j) are connected and $Y_{ij} = 0$ otherwise.
The basic assumption of these models is that the structure in an observed graph y can be explained by a statistics $s(y)$ depending on the observed network and unit attributes. This way, it is possible to describe any kind of dependence between the dyadic variables:

$$P(Y = y|\theta) = \frac{\exp(\theta^T s(y))}{c(\theta)}$$

where θ is a vector of model parameters associated with $s(y)$ and $c(\theta)$ is a normalising constant.
See more in Robins (2011).

Actor-based model for network dynamics

- Actor-based model can be used for longitudinal network data (Snijders, 2001, 2005; Snijders et al., 2007, 2010).
- The model is defined as a continuous-time Markov chains model, observed possibly only at a few discrete time points or time intervals, where tie changes are modeled as choices by units (actors).
- The probability model for tie changes is based on linear predictors similar to generalized linear models.

The model is implemented in the program SIENA.
See more in Snijders (2011).

Estimated parameters for the six scientific fields in Slovenia (green estimates are not statistically significant)

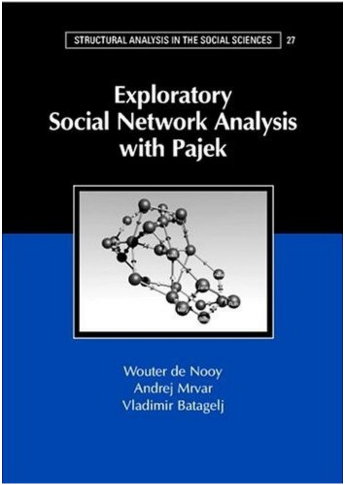
	1 Nat	2 Eng	3 Med	4 Bio	5 Soc	6 Hum
rate 1	22.06	18.49	45.01	29.30	32.65	14.84
rate 2	27.21	26.03	54.95	32.98	35.26	16.58
degree (density)	-2.36	-2.55	-2.11	-1.66	-2.40	-3.45
transitivity	0.46	0.71	0.35	0.37	0.45	1.73
same research group	1.54	2.02	1.26	0.92	1.49	2.29
degree of alter	-0.02	-0.06	-0.02	-0.02	-0.05	-0.10
degree out	0.17	0.21	0.17	0.13	-0.03	-0.15
excellence	-0.12	−0.01	−0.01	−0.03	0.53	0.54
age	0.01	0.01	0.02	0.02	0.01	0.01
age similarity	0.11	−0.07	0.02	0.24	0.05	−0.33
PhD (yes)	0.99	1.02	0.71	0.48	0.92	0.50
Gender (male)	0.10	0.17	0.17	0.17	-0.24	0.19

Software for SNA

UCINET, NetDraw	Pajek	Netminer
Visone	SNA/R	StOCNET
Negopy	InFlow	GUESS
NetworkX	prefuse	JUNG
BGL/Python		

An overview of the software for SNA is in Huisman and van Duijn (2011).

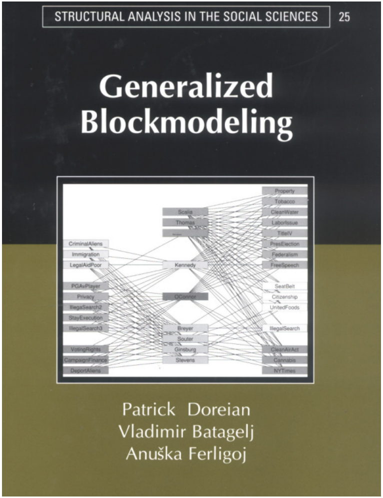
Pajek

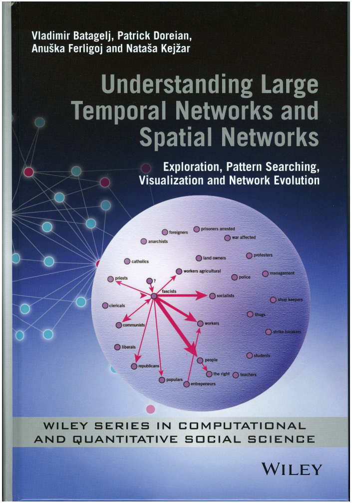


An introduction to social network analysis with `Pajek` is available in the book ESNA (de Nooy, Mrvar, Batagelj 2005; second edition 2011).

`Pajek` – program for analysis and visualization of large networks is freely available, for noncommercial use, at its web site.

`http://pajek.imfm.si/`


STRUCTURAL ANALYSIS IN THE SOCIAL SCIENCES
25
Generalized Blockmodeling
Patrick Doreian
Vladimir Batagelj
Anuška Ferligoj


Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj and Nataša Kejžar
Understanding Large Temporal Networks and Spatial Networks
Exploration, Pattern Searching, Visualization and Network Evolution
WILEY SERIES IN COMPUTATIONAL AND QUANTITATIVE SOCIAL SCIENCE
WILEY