Network Analysis 2
Statistical Approaches and Modeling

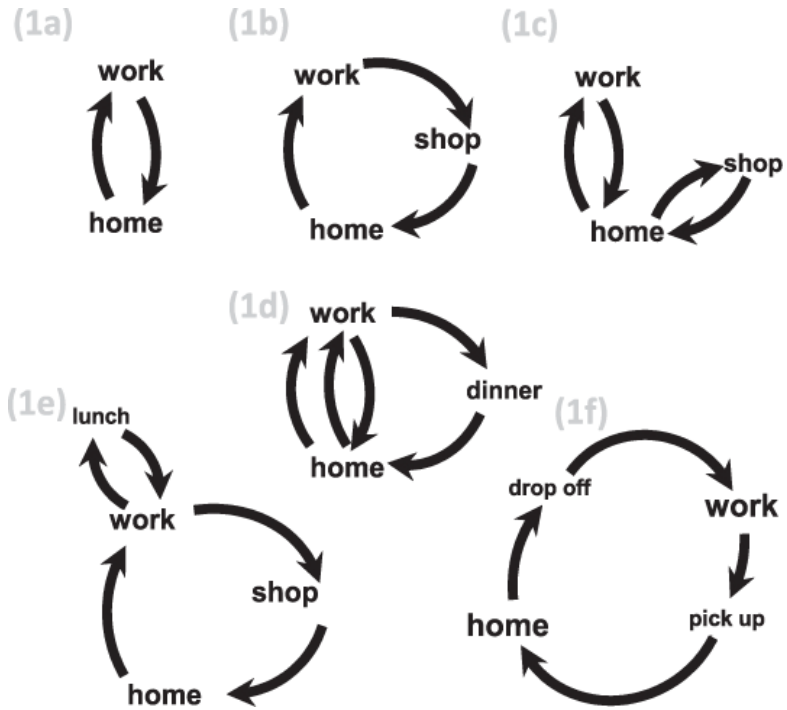
Patterns

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, and NRU HSE Moscow

XIII Summer School of the ANR-Lab
Network Analysis and Contemporary Decision Sciences
International Laboratory for Applied Network Research
NRU HSE, Moscow, August 2022

Outline



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Current version of slides (August 22, 2022 at 15 : 02): slides PDF

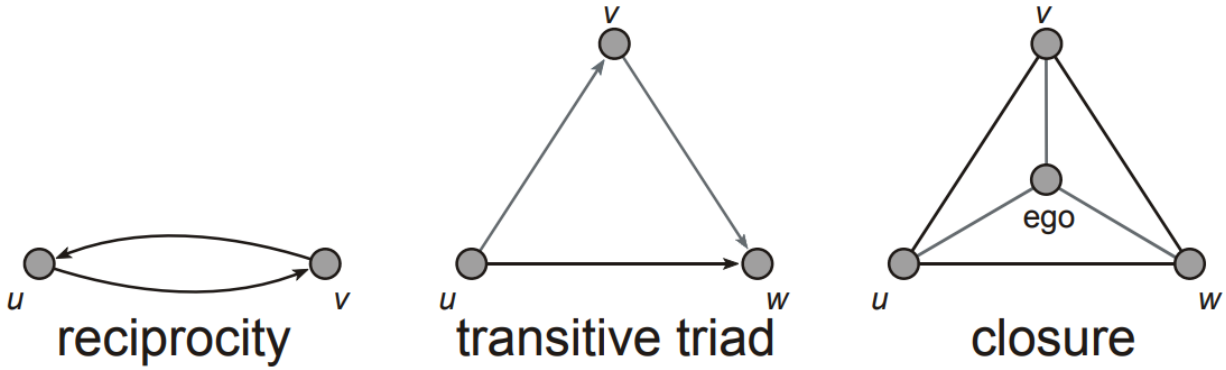
Subgroups

Where do cohesive subgroups come from and what do they do?

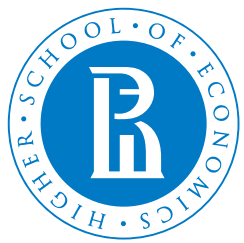
The first and most general behavioural hypothesis merely states that similar people tend to interact more easily and people who interact tend to become or perceive themselves as more similar provided that the interaction is characterized as positive, friendly, and so on. In SNA, this tendency is mainly known as *homophily*, a concept coined by Paul F. Lazarsfeld and Robert K. Merton, but it is known under other names in several scientific disciplines, e.g., the phenomenon of attribution and affect control in social-psychology, assortative or selective mixing in epidemiology and ecology, and assortative mating in genetics.

If we concentrate on the graph theoretical aspects of this behavioural hypothesis, that is, the structure of ties, and take the (dis)similarities among the actors for granted, we find several characteristics of local structures that measure cohesive subgroup formation. At the level of a pair of actors, *reciprocity* of ties in directed networks (arcs) signals subgroup formation: both actors are hypothesized to choose each other when they are similar.

Subgroups



At the level of the triple, *transitivity* results from tendencies toward cohesion. If actor *u* establishes a tie with actor *v* because they are similar, and actor *v* establishes a tie to actor *w* for the same reason, actors *u* and *w* are also similar, so actor *u* is expected to establish a tie with *w* as well, creating a so-called transitive triad. Stated differently, the path from *u* via *v* to *w* is closed by an arc from *u* to *w*. In general, the *closure* of paths or semipaths both in directed and undirected networks signals cohesive subgroup formation at the local level. Closure within an ego-network may be calculated as the percentage of all possible ties among a vertex' neighbours that are present, which is one of the definitions of the clustering coefficient. The concept of closure can be extended beyond a vertex' immediate neighbors.

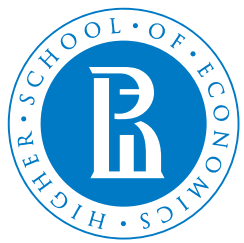


Subgroups

If we include measured attributes of the vertices in our indicators of cohesive subgroup formation, we can calculate homophily quite simply as the probability or ratio of ties between vertices that share a particular characteristic to ties between vertices that do not. Extending this idea to the ego-network, the homogeneity of actors involved in an ego-network may be taken as a measure of tendencies toward homophily.

For qualitative attributes of the actors, *Blau's index* of variability or heterogeneity, $B = 1 - \sum p_i^2$, can be used, where p_i is the proportion of group members in a particular category. It is conceptually related to the Herfindahl-Hirschman Index in economics measuring the extent of monopoly within an industry. It is interesting to note that Blau's theory hypothesizes that heterogeneity rather than homogeneity of actors within a group enhances the operation and efficiency of the group. If improving group *efficiency* is the aim of actors, we would have to use a behavioural hypothesis that is quite the opposite of the homophily hypothesis.

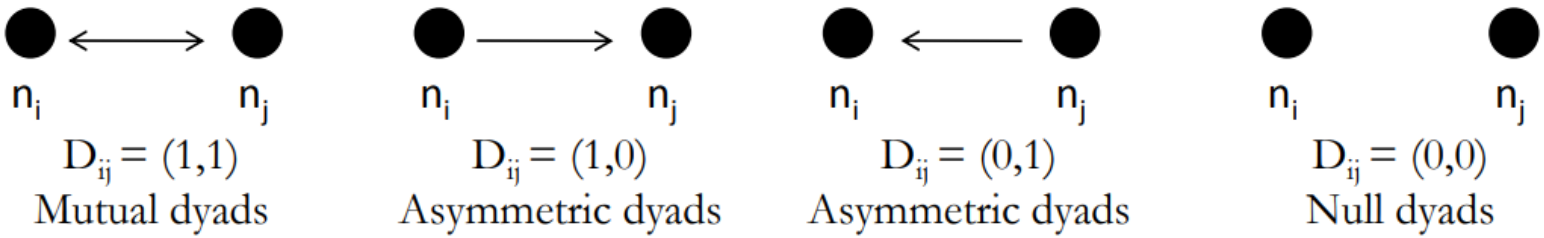
WP: diversity



Subgroups

In addition to homophily, there is a second behavioural hypothesis related to cohesion in SNA. This hypothesis is based on the idea that social action is embedded in networks. Named after the sociologist Georg Simmel, *Simmelian ties* are ties that are embedded in other ties, e.g., business ties are embedded in family ties, or in complete triads and cliques. Embedded ties are hypothesized to enforce group norms and enhance trust, hence pressure people into the same behaviour because there are parallel ties or because the two actors involved in a tie share common neighbours who supervise their behaviour. In the Florentine families example, we see that eight out of fifteen business ties are backed up by marriage alliances.

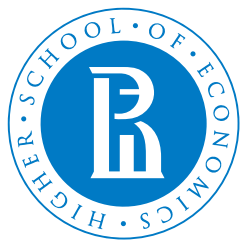
Dyads



Three dyadic isomorphism classes for directed graphs:

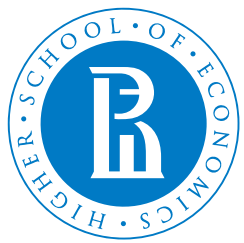
- **null** dyads have no arcs
- **asymmetric** dyad has an arc between the two nodes going in one direction or the other, but not both.
- **mutual** dyads have two arcs between the nodes, one going in one direction, and the other going in the opposite direction.

$$\text{reciprocity} = \frac{\#\text{ mutual}}{\#\text{ non-null}}$$



Valued reciprocity

Traditional studies defined reciprocity in a very simple, but fundamentally limited, way. A dyad was reciprocal if both partners nominated one another as friends, or, in the tradition of “balance theory”, if it was found that the relationship had the same valence (positive or negative) for both participants. Dyads were viewed as nonreciprocal either when one partner reported considering the other one a friend or a close associate and the other did not, or if one partner displayed positive sentiments towards a partner who felt negatively towards him or her. The fundamental hypothesis of balance concerned a dynamic prediction: over time ties that were imbalanced were expected either to become balanced or to dissolve.



Valued reciprocity

This definition of "reciprocity" fit very well with the representation of social networks in early graph theory as consisting of binary (1,0) edges connecting two nodes. Analysts can then establish the level of reciprocity in the network via the so-called *dyadic census*, otherwise known as the UMAN classification. The phenomenon of dyadic reciprocity at the level of the whole network has been studied by comparing the relative distribution of asymmetric and mutual dyads in a graph. Non-reciprocity is high if the proportion of asymmetric dyads is larger than would be obtained by chance in a graph with similar topological properties (for instance a graph with the same number of vertices and edges).

The idea here is that low status actors direct nominations towards high-status actors with those nominations unlikely to be returned; reciprocal nominations, on the other hand, should be more common among actors of comparable rank. If one actor initiates all directed communication attempts while the other actors initiates none, then reciprocity is not defined ($R_{uv} = \infty$), which is consistent with the intuition that there can be no definition of reciprocity when there is no actual two-way relationship to speak of.

Valued reciprocity

In spite of its utility, the binary classification of dyads into three types misses one of the most important features of a dyadic relationship: the relative frequency of contact between the two partners.

A more empirically accurate definition of reciprocity can only be obtained in the context of a weighted graph [14] . Consistent with the notion of reciprocity as balance, this index should have the following properties:

1. it should be at a minimum (indicating reciprocity) when the weight of the directed arc going from vertex u to vertex v approaches the weight of the directed arc going from vertex v to vertex u.
2. it should increase monotonically with the weight difference between the two directed arcs.
3. it should normalize the weight difference to adjust for the fact that some persons are simply more or less communicative than others (they contact all of their partners more or less frequently).
4. the index should be the same irrespective of directionality ($R_{uv} = R_{vu}$).

Valued reciprocity

One index that satisfies these conditions is:

$$R_{uv} = |\ln p_{uv} - \ln p_{vu}|$$

With $p_{uv} = w_{uv}/w_{u+}$ where w_{uv} is the weight corresponding to the arc (u, v), and w_{u+} is the strength of the vertex u as given by $w_{u+} = \sum_{v \in N(u)} w_{uv}$.

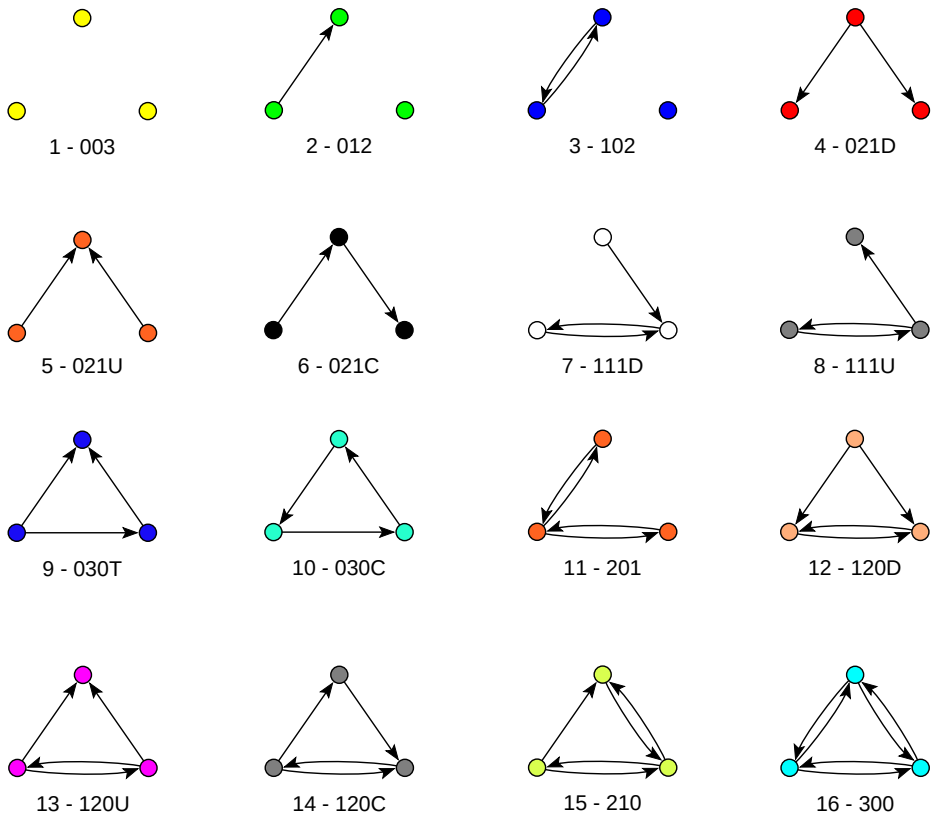
A substantive interpretation of a reciprocity index based on the ratio of normalized weights is that a dyad is reciprocal when two persons have the same probability of communicating with one another.

$$R_{uv} = |\ln \frac{w_{uv} w_{v+}}{w_{vu} w_{u+}}|$$

The first idealized condition that we can consider is an equidispersion regime. Under this condition, persons distribute their communicative activity equally across partners – under this regime the expected directed weights are given by: $\overline{w}_{uv} = w_{u+}/\text{outdeg}(u)$ or $\overline{p}_{uv} = 1/\text{outdeg}(u)$. Finally, the reciprocity equation simplifies to:

$$\overline{R}_{uv} = |\ln \text{outdeg}(u) - \ln \text{outdeg}(v)|.$$

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 nodes. There are 16 nonisomorphic (types of) triads. They can be partitioned into three basic types:

- the *null* triad 003;
- *dyadic* triads 012 and 102; and
- *connected* triads: 111D, 201, 210, 300, 021D, 111U, 120D, 021U, 030T, 120U, 021C, 030C and 120C.

`Network/Info/Triadic Census`

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

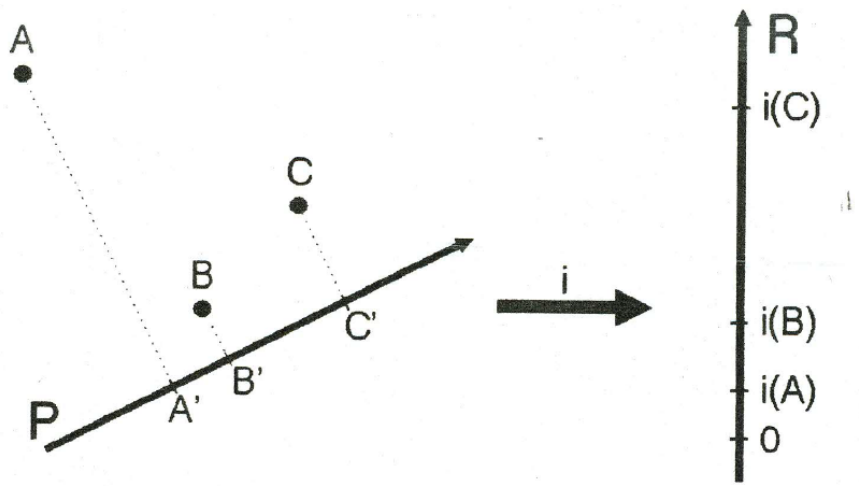
Several properties of a graph can be expressed in terms of its *triadic spectrum* – distribution of all its triads. It also provides ingredients for p^* network models.

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.

Structural indices

Given a set of units (descriptions of objects) $\mathcal{E}$ and their structural property P (e.g. size, connectivity, reciprocity, etc.) which is described by an intuitive or empirical relation **P**

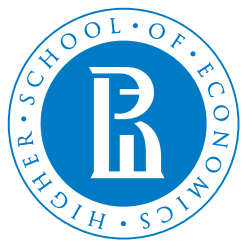
$$X\mathbf{P}Y \equiv \text{ unit } X \text{ is less-}P \text{ than unit } Y$$



We say that the mapping $i : \mathcal{E} \rightarrow \mathbb{R}$ is an *index* measuring the property P if it satisfies the condition

$$X\mathbf{P}Y \Leftrightarrow i(X) < i(Y)$$

and possibly some other conditions reflecting transformations on units.

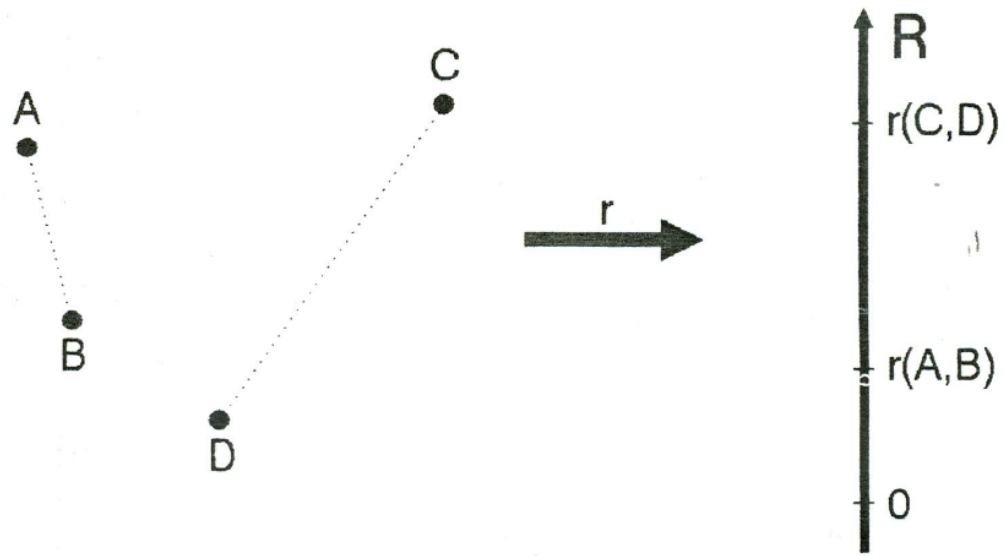


Structural indices

Each index induces a ”dimension” in the ”space” of units. From figure we see that unit *B* is closer to *A* than to *C with respect to the property P*; but in the ”space” the unit *B* is closer to *C* than to *A*.

With few exceptions the inverse way to define an index is usually used. This leads to the *interpretation problem*: For a given index *i* what is the meaning of the corresponding property *P*?

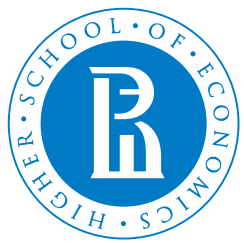
Dissimilarities



Let $\mathcal{E}$ be a set of units. Quantitatively we describe the *resemblance* (association, similarity) between units by a function (*resemblance measure*)

$$r : (X, Y) \mapsto \mathbb{R}$$

Many examples of resemblances for different types of units can be found in any book on data analysis and related topics [7].



Dissimilarities

For r to be a resemblance, we require that it is *symmetric*:

P1. $\forall X, Y \in \mathcal{E} : \; r(X, Y) = r(Y, X)$

and that it has either the property:

P2.a $\forall X, Y \in \mathcal{E} : \; r(X, X) \leq r(X, Y)$,

or the property:

P2.b $\forall X, Y \in \mathcal{E} : \; r(X, X) \geq r(X, Y)$.

A resemblance which satisfies condition P2.a is called *forward* (straight) and denoted by d; it is called *backward* (reverse) and denoted by s if it satisfies condition P2.b.

Dissimilarities

In the set of unordered pairs of units $\mathcal{E}_2 = \{[X, Y] : X, Y \in \mathcal{E}\}$, $[X, Y] = [Y, X]$, a resemblance r induces the ordering $\ll_r$ as follows:

$$[X, Y] \ll_r [U, V] \equiv r(X, Y) < r(U, V)$$

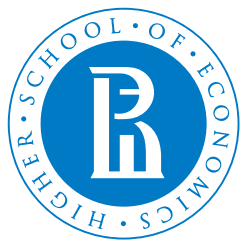
The relation $\ll_r$ is a strict partial order. Resemblances r and s are *(order) equivalent*, $r \cong s$, iff: $\ll_r = \ll_s$ or $\ll_r = \ll_s^{-1}$. $\cong$ is an equivalence relation. Also:

Theorem: Let $f : r(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}$ be a strictly increasing/decreasing function and r a resemblance. Then

$$s(X, Y) = f(r(X, Y)) \qquad \text{for all } X, Y \in \mathcal{E}$$

is also a resemblance and $s \cong r$.

An important consequence of this theorem is that every backward resemblance measure s can always be transformed by $d(X, Y) = -s(X, Y)$ into an order equivalent forward resemblance measure d. Therefore in the following we can limit our discussion to forward resemblances.



Dissimilarities

Forward resemblances usually have the property:

P3.a $\quad \exists\, r^* \in \mathbb{R}\, \forall\, X \in \mathcal{E} :\ r(X,X) = r^*.$

In this case we can define a new resemblance d:

$$d(X,Y) = r(X,Y) - r^*$$

which is order equivalent to r and has the properties:

R1. $\forall\, X, Y \in \mathcal{E} :\quad d(X,Y) \geq 0;$
R2. $\forall\, X \in \mathcal{E} :\quad d(X,X) = 0;$
R3. $\forall\, X, Y \in \mathcal{E} :\quad d(X,Y) = d(Y,X).$

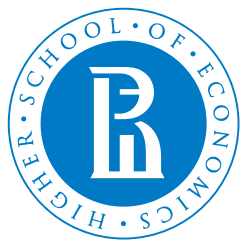
A resemblance d satisfying properties R1, R2 and R3 is called a *dissimilarity*. Many data analysis algorithms deal with dissimilarities.

Dissimilarities

For some dissimilarities, additional properties hold:

R4. *evenness*: $d(X,Y) = 0 \Rightarrow \forall Z: \; d(X,Z) = d(Y,Z)$;
R5. *definiteness*: $d(X,Y) = 0 \Rightarrow X = Y$;
R6. *triangle inequality*: $d(X,Y) \leq d(X,Z) + d(Z,Y)$;
R7. *ultrametric inequality*: $d(X,Y) \leq \max(d(X,Z), d(Z,Y))$;
R8. *Buneman's inequality* or *four-points condition*:
$d(X,Y) + d(U,V) \leq \max(d(X,U) + d(Y,V), d(X,V) + d(Y,U))$;
R9. *translation invariance*: Let $(\mathcal{E}, +)$ be a group
$d(X,Y) = d(X+Z, Y+Z)$.

These properties are related in the following way:
R7 $\Rightarrow$ R6 $\Rightarrow$ R4 $\Leftarrow$ R5 and R8 $\Rightarrow$ R6.
Dissimilarity d which has also the properties R5 and R6 is called a *distance*. Monotone hierarchical clustering algorithms transform dissimilarities into ultrametric dissimilarities. Dissimilarities satisfying Buneman's inequality are *tree distances* – distances between units are the shortest path lengths in some tree.



Dissimilarities

Dissimilarities usually take values in the interval $[0, 1]$ or in the interval $[0, \infty]$. They can be transformed one into the other by mappings:

$$\frac{d}{1-d} : [0,1] \rightarrow [0,\infty] \quad \text{and} \quad \frac{d}{1+d} : [0,\infty] \rightarrow [0,1],$$

or in the case $d_{max} < \infty$ by

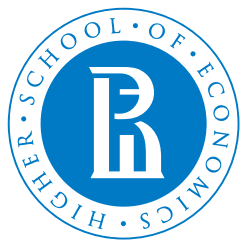
$$\frac{d}{d_{max}} : [0, d_{max}] \rightarrow [0,1].$$

To transform distance into distance we often use the mappings:

$$\log(1+d), \quad \min(1,d) \quad \textit{and} \quad d^r,\ 0 < r < 1.$$

Not all resemblances are dissimilarities. For example, the correlation coefficient has the interval $[1, -1]$ as its range. We can transform it to the interval $[0, 1]$ by mappings:

$$\tfrac{1}{2}(1-d), \quad \sqrt{1-d^2}, \quad 1-|d|, \ \ldots$$



Dissimilarities

When applying these transformations to a measure d we wish that the nice properties of d were preserved. In this respect the following theorems should be mentioned:

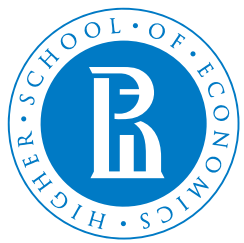
Theorem: Let d be a dissimilarity on $\mathcal{E}$ and let a mapping $f: d(\mathcal{E} \times \mathcal{E}) \rightarrow \mathbb{R}_0^+$ has the property $f(0) = 0$, then $d'(X, Y) = f(d(X, Y))$ is also a dissimilarity.

Theorem: Let d be a distance on $\mathcal{E}$ and let the mapping $f: d(\mathcal{E} \times \mathcal{E}) \rightarrow \mathbb{R}$ has the properties:

(a) $f(x) = 0 \Leftrightarrow x = 0$,
(b) $x < y \Rightarrow f(x) < f(y)$,
(c) $f(x + y) \leq f(x) + f(y)$,

then $d'(X, Y) = f(d(X, Y))$ is also a distance and $d' \cong d$.

All concave functions have also the sub-additivity property (c).



Dissimilarities

The following concave functions satisfy the last theorem:

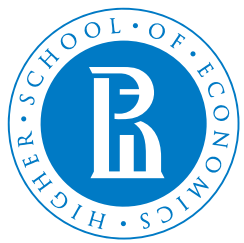
(a)	$f(x) = \alpha x,\ \alpha > 0,$	(b)	$f(x) = \log(1+x),\ x \geq 0,$
(c)	$f(x) = \frac{x}{1+x},\ x \geq 0,$	(d)	$f(x) = \min(1, x),$
(e)	$f(x) = x^{\alpha},\ 0 < \alpha \leq 1,$	(f)	$f(x) = \arcsin x,\ 0 \leq x \leq 1.$

Theorem: Let $d : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ has the property Ri, $i = 1, ..., 7$, then $f(d)$, $f \in$ (a)-(f) also has this property.

Some operations preserve properties Ri, $i = 1, ..., 7$:

Theorem: Let $d_1 : \mathcal{E}_1 \times \mathcal{E}_1 \to \mathbb{R}$ and $d_2 : \mathcal{E}_2 \times \mathcal{E}_2 \to \mathbb{R}$ have property Ri, then $(d_1 +_p d_2)((X_1, X_2), (Y_1, Y_2)) = \sqrt[p]{a \cdot d_1(X_1, Y_1)^p + b \cdot d_2(X_2, Y_2)^p}$, $a, b > 0$ also has property Ri, $i = 1, \ldots, 5, 7$ for $p > 0$; and also has property R6 for $p \geq 1$ over $\mathcal{E}_1 \times \mathcal{E}_2$.

The theorem holds also in the case when $\mathcal{E}_1 = \mathcal{E}_2$ and $X_1 = X_2,\ Y_1 = Y_2$.



Dissimilarities

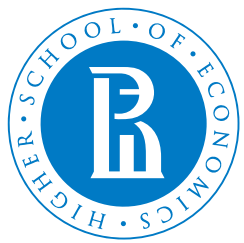
How to define dissimilarities between structured objects?

Given an index i, we can always define a dissimilarity

$$d_i(X, Y) = |i(X) - i(Y)|$$

It has also the property R6.

The main drawback of so defined dissimilarity is its unidimensionality – it compares units only with respect to selected property measured by the index i. This deficiency can be overcome by combining dissimilarities based on different indices using the previous theorem. Coefficients a and b can be used to "tune" the influence of each property.



Dissimilarities

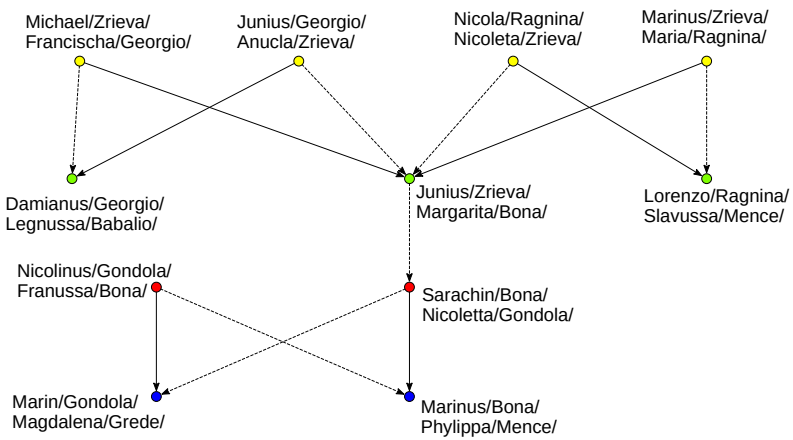
To a structured object X we assign its *description* – a vector consisting of values of selected structural indices. The dissimilarity between objects is then defined as some (standard) dissimilarity between their descriptions. The selection of this dissimilarity depends on our problem and types of scales in which the properties are measured. In the case of different types of scales Gower's dissimilarity is often used gower.dist.

The main problems in this approach are:

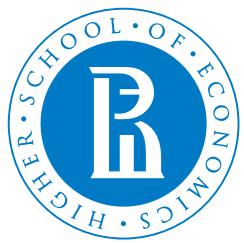
- **completness:** do we consider all (for our purposes) important properties of units?
- **"correlations":** are there some relations among the selected properties/indices?
- **weighting:** is the right level of importance given to all indices combined in the dissimilarity?

Patterns

If a selected *pattern* or fregment 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.



Three connected relinking marriages in the genealogy (represented as a *p*-graph) of ragusan noble families. A solid arc indicates the _ *is a son of* _ relation, and a dotted arc indicates the _ *is a daughter of* _ relation. In all three patterns a brother and a sister from one family found their partners in the same other family.



Patterns

The basic question is what is considered as an appearance of a pattern in a network. Is this an induced subnetwork isomorphic to a given pattern; or the induced subnetwork has to contain a given pattern as a subnetwork?

Additionally, three specific concepts of sub-graph appearance have been proposed.

- **A:** considers all matches of a graph in original network;
- **B:** is searching for the maximum number of link-disjoint instances of a given graph in original network;
- **C:** considers matches with disjoint links and nodes.

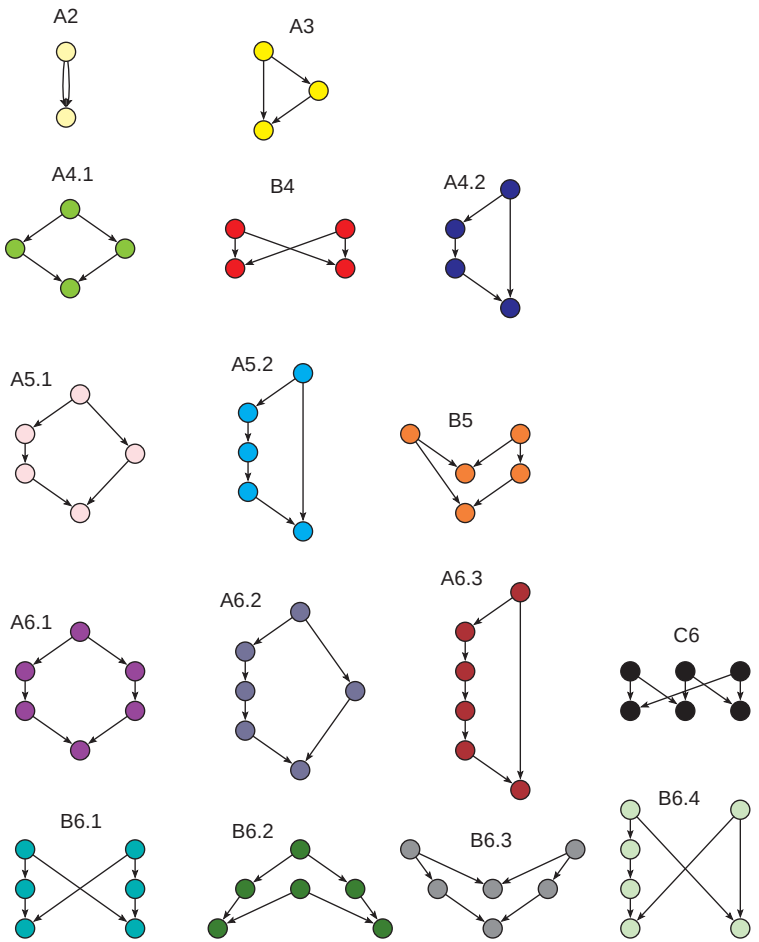
...Pattern searching

To speed up the search or to consider some additional properties of the pattern, a user can set some additional options:

- nodes in network should match with nodes in pattern in some nominal, ordinal or numerical property (for example, type of atom in molecula);
- values of edges must match (for example, edges representing male/female links in the case of *p-graphs*);
- the first node in the pattern can be selected only from a given subset of nodes in the network.

`Networks/Fragment (First in Second)`

Relinking patterns in *p*-graphs

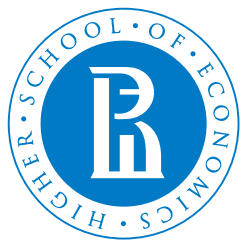


frag16.paj

All possible relinking marriages in *p*-graphs with 2 to 6 nodes. Patterns are labeled as follows:

- first character – number of first nodes: A – single, B – two, C – three.
- second character: number of nodes in pattern (2, 3, 4, 5, or 6).
- last character: identifier (if the two first characters are identical).

Patterns denoted by A are exactly the blood marriages. In every pattern the number of first nodes is equal to the number of last nodes.

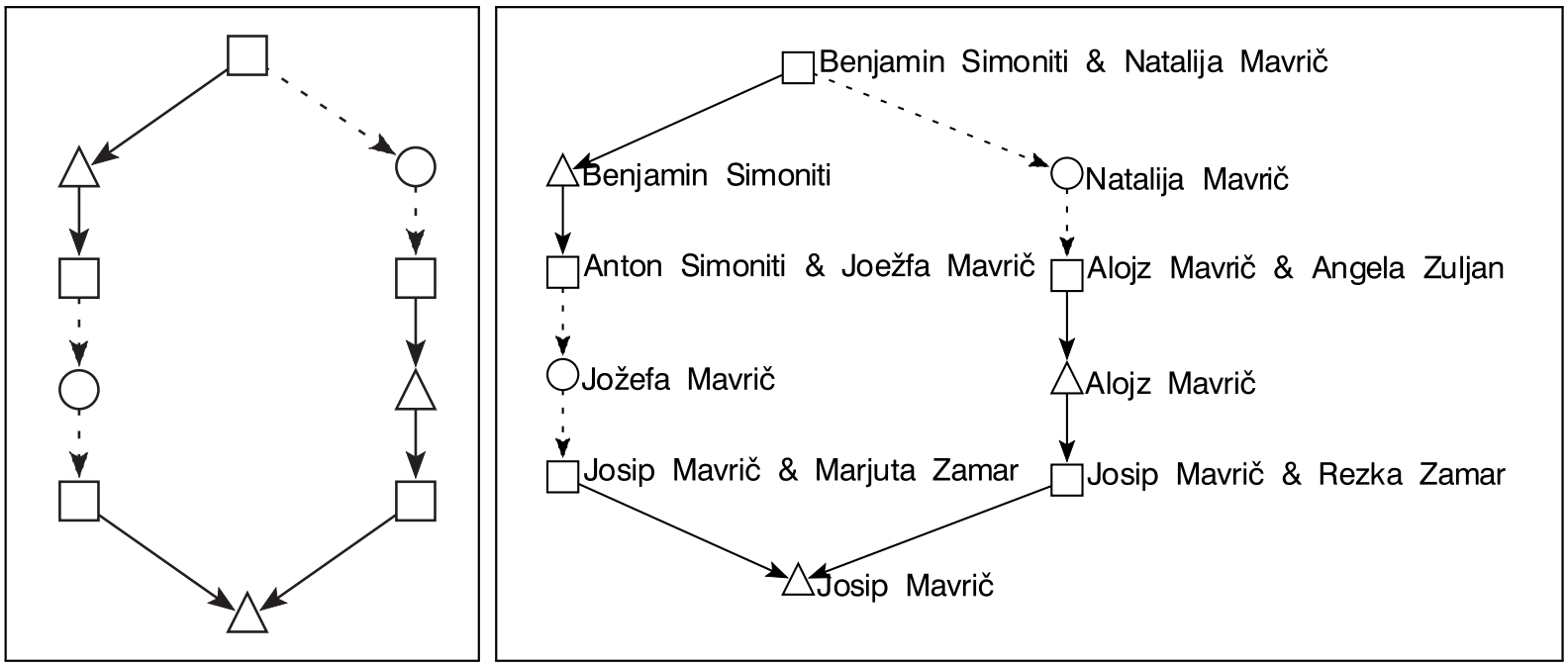


Frequencies normalized with number of couples in *p*-graph $\times$ 1000

pattern	Loka	Silba	Ragusa	Turcs	Royal
A2	0.07	0.00	0.00	0.00	0.00
A3	0.07	0.00	0.00	0.00	2.64
A4.1	0.85	2.26	1.50	**159.71**	18.45
B4	3.82	11.28	10.49	**98.28**	6.15
A4.2	0.00	0.00	0.00	0.00	0.00
A5.1	0.64	3.16	2.00	36.86	11.42
A5.2	0.00	0.00	0.00	0.00	0.00
B5	1.34	4.96	23.48	46.68	7.03
A6.1	1.98	12.63	1.00	**169.53**	11.42
A6.2	0.00	0.90	0.00	0.00	0.88
A6.3	0.00	0.00	0.00	0.00	0.00
C6	0.71	5.41	9.49	36.86	4.39
B6.1	0.00	0.45	1.00	0.00	0.00
B6.2	1.91	17.59	31.47	**130.22**	10.54
B6.3	3.32	13.53	40.96	**113.02**	11.42
B6.4	0.00	0.00	2.50	7.37	0.00
Sum	14.70	72.17	123.88	798.53	84.36

Most of the relinking marriages happened in the genealogy of Turkish nomads; the second is Ragusa while in other genealogies they are much less frequent.

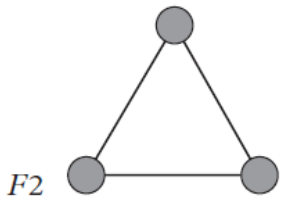
Bipartite *p*-graphs: Marriage among half-cousins



Pattern counting using matrices

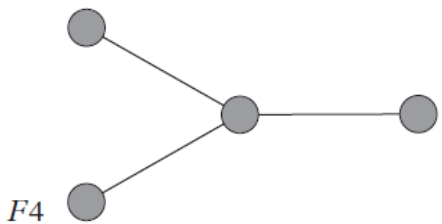
In the book [8] a long list of formulae for counting small subgraphs is given.

$$|F_1| = \frac{1}{2}\sum_i k_i(k_i-1)$$



$$|F_2| = \frac{1}{6}\mathrm{tr}(A^3)$$

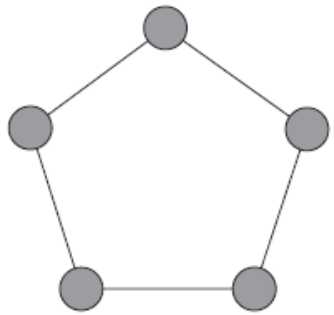
$$|F_3| = \sum_{(i,j)\in E}(k_i-1)(k_j-1) - 3|F_2|$$



$$|F_4| = \frac{1}{6}\sum_i k_i(k_i-1)(k_i-2)$$

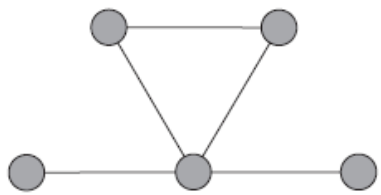
In physics they denote degrees by k.

Pattern counting using matrices



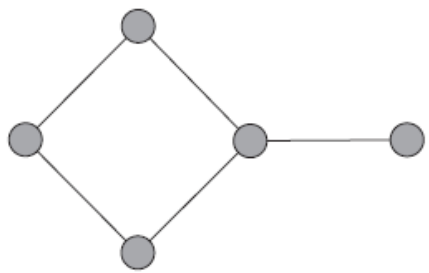
F8

$$|F_8| = \frac{1}{10}\left(\mathrm{tr}(A^5) - 30|F_2| - 10|F_6|\right)$$



F9

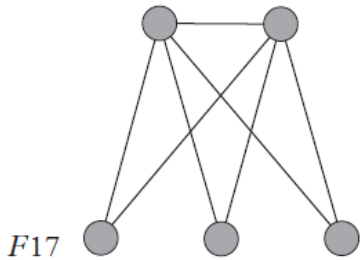
$$|F_9| = \frac{1}{2}\sum_{k_i \geq 4} t_i(k_i - 2)(k_i - 3)$$



F10

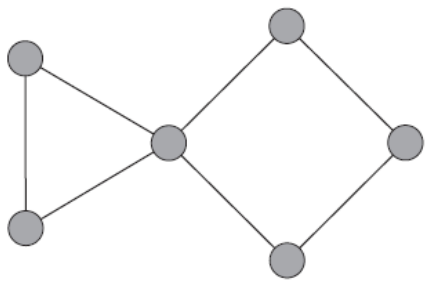
$$|F_{10}| = \frac{1}{2}\sum_i (k_i - 2) \times \sum_{i,j} \binom{(A^2)_{ij}}{2} - 2|F_7|$$

Pattern counting using matrices



$F17$

$$|F_{17}| = \sum_{(i,j)\in E} \binom{(A^2)_{ij}}{3}$$



$F18$

$$|F_{18}| = \sum_i t_i \cdot \sum_{i\neq j} \binom{(A^2)_{ij}}{2} - 6|F_7| - 2|F_{14}| - 6|F_{17}|$$

Motifs

Network motifs are sub-graphs that repeat themselves in a specific network or even among various networks. Each of these sub-graphs, defined by a particular pattern of interactions between vertices, may reflect a framework in which particular functions are achieved efficiently. Indeed, motifs are of notable importance largely because they may reflect functional properties.

Motifs

When $H \subseteq G$ and there exists an isomorphism between the sub-graph H and a graph F, this mapping represents an appearance of F in G. The number of appearances of graph F in G is called the frequency of F in G. It is denoted by $f_G(H)$.

A graph F is called recurrent (or frequent) in G, when its frequency $f_G(F)$ is above a predefined threshold or cut-off value. There is a class $\Omega(G)$ of random graphs corresponding to the null-model associated to G. We should choose N random graphs uniformly from $\Omega(G)$ and calculate the frequency for a particular frequent sub-graph F in G. If the frequency of F in G is higher than its arithmetic mean frequency in N random graphs R_i, where $1 \leq i \leq N$, we call this recurrent pattern significant and hence treat F as a network *motif* for G.

Motifs

For a small graph F, the network G and a set of randomized networks $\mathbf{R}(G) \subseteq \Omega(G)$, where $|\mathbf{R}(G)| = N$, the Z-Score that has been defined by the following formula:

$$Z(F) = \frac{f_G(F) - \mu_R(F)}{\sigma_R(F)}$$

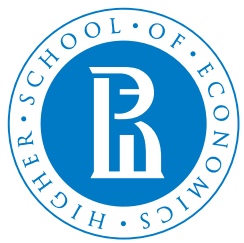
where $\mu_R(F)$ and $\sigma_R(F)$ stand for mean and standard deviation frequency in set $R(G)$, respectively. The larger the $Z(F)$, the more significant is the sub-graph F as a motif.

Motifs

Alternatively, another measurement in statistical hypothesis testing that can be considered in motif detection is the *p*-value, given as the probability of $f_R(F) \geq f_G(F)$ (as its null-hypothesis), where $f_R(F)$ indicates the frequency of F in a randomized network. A sub-graph with *p*-value less than a threshold (commonly 0.01 or 0.05) will be treated as a significant pattern. The *p*-value is defined as

$$P(F) = \frac{1}{N} \sum_{i=1}^{N} \delta(c(i)) \quad \text{with} \quad c(i) \equiv f_R^i(F) \geq f_G(F)$$

Where N indicates number of randomized networks, i is defined over a class of randomized networks and the Kronecker delta function $\delta(c(i))$ is one if the condition $c(i)$ holds.



Motifs

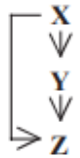
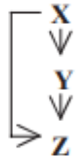
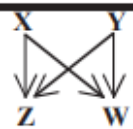
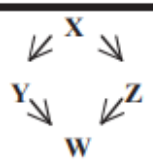
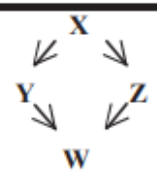
The concentration of a particular n-size sub-graph F in network G refers to the ratio of the sub-graph appearance in the network to the total n-size non-isomorphic sub-graphs' frequencies, which is formulated by

$$C_G(F) = \frac{f_G(F)}{\sum_i f_G(G_i)}$$

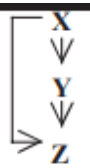
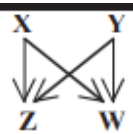
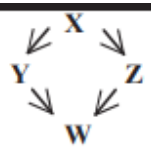
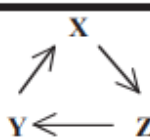
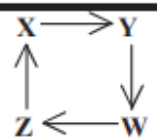
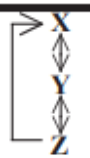
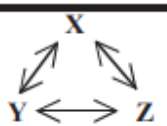
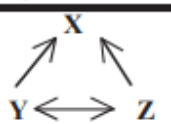
where index i is defined over the set of all non-isomorphic n-size graphs.

Another statistical measurement is defined for evaluating network motifs, but it is rarely used in known algorithms. This measurement is introduced by Picard et al. in 2008 and used the Poisson distribution, rather than the Gaussian normal distribution that is implicitly being used above.

Motifs

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

Motifs

					Feed-forward loop			Bi-fan			Bi-parallel
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					Three-node feedback loop			Bi-fan			Four-node feedback loop
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					Feedback with two mutual dyads			Fully connected triad			Uplinked mutual dyad
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

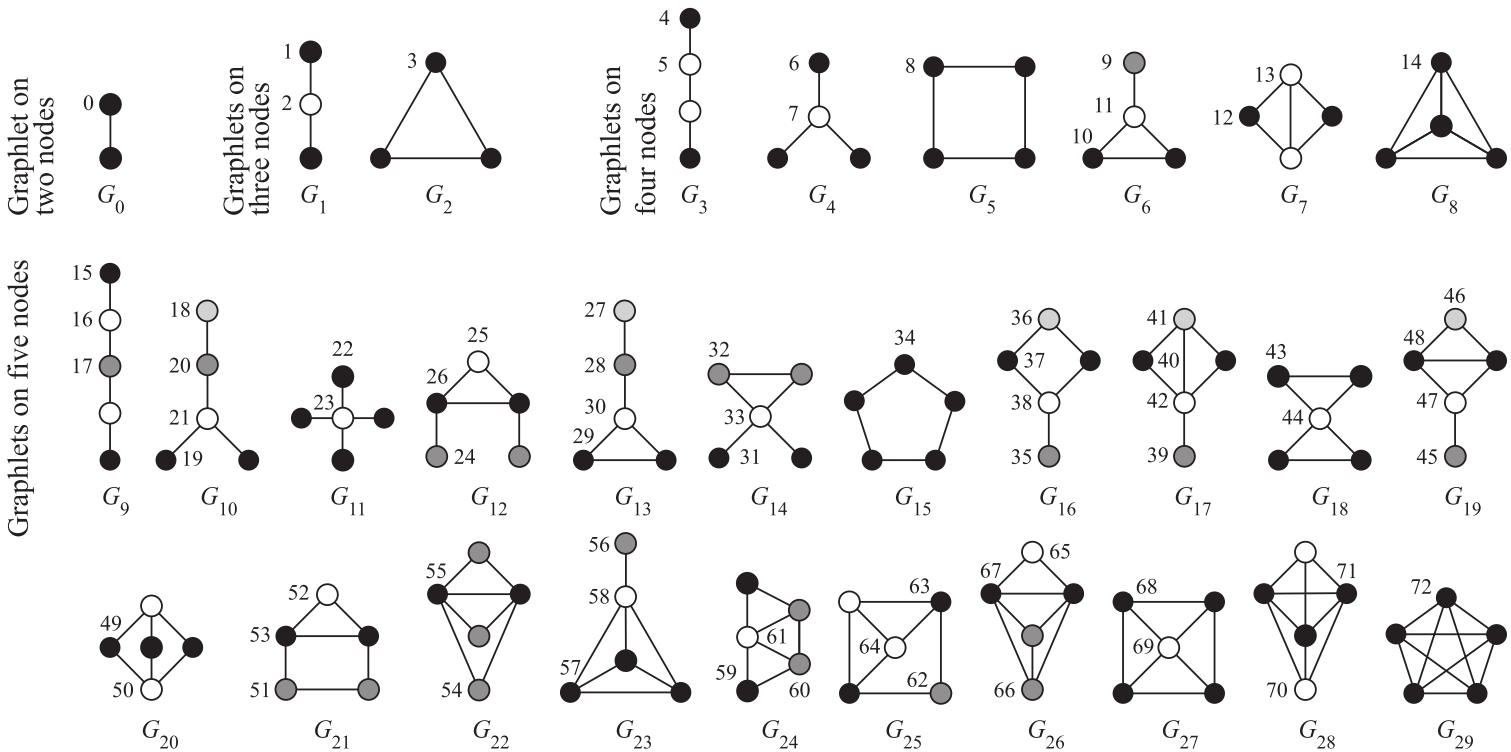
Many studies/applications in biology.
No attention yet to “forbiden” motifs (see triads).
R: igraph::motifs

Graphlets

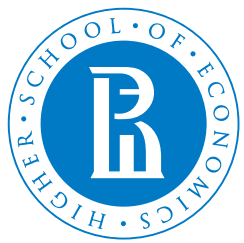
Graphlets are small connected non-isomorphic induced subgraphs of a large network. Graphlets differ from network motifs, since they must be induced subgraphs, whereas motifs are partial subgraphs. An induced subgraph must contain all edges between its nodes that are present in the large network, while a partial subgraph may contain only some of these edges.

Graphlets were first introduced by Nataša Pržulj.

Graphlets with 2–5 nodes and automorphism orbits



Nodes of the same color belong to the same orbit within that graphlet.

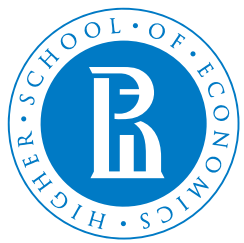


Relative graphlet frequency distance

RGF-distance compares the frequencies of the appearance of all 3-5-node graphlets in two networks.
Let Ni(G) be the number of graphlets of type $i \in \{0, \ldots, 29\}$ in network G, and let $T(G) = \sum_{i=0}^{29} N_i(G)$ be the total number of graphlets of G. The "similarity" between two graphs should be independent of the total number of nodes or edges, and should depend only upon the differences between relative frequencies of graphlets. Thus, relative graphlet frequency distance $D(G, H)$ between two graphs G and H is defined as:

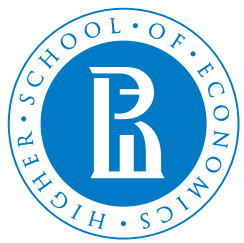
$$D(G, H) = \sum_{i=0}^{29} |F_i(G) - F_i(H)|$$

where $F_i(G) = -\log(N_i(G)/T(G))$. The logarithm of the graphlet frequency is used because frequencies of different graphlets can differ by several orders of magnitude and the distance measure should not be entirely dominated by the most frequent graphlets.



Graphlet degree distribution agreement

The degree distribution measures the number of nodes of degree k in graph G, i.e., the number of nodes "touching" k edges, for each value of k. Note that an edge is the only graphlet with two nodes. GDDs generalize the degree distribution to other graphlets: they measure for each 2-5-node graphlet G_i, $i = 0, 1, ..., 29$, such as a triangle or a square, the number of nodes "touching" k graphlets G_i at a particular node. A node at which a graphlet is "touched" is topologically relevant, since it allows us to distinguish between nodes "touching", for example, a three node path at an end node or at the middle node. This is summarized by automorphism orbits (or just orbits, for brevity): by taking into account the "symmetries" between nodes of a graphlet, there are 73 different orbits across all 2-5-node graphlets.



Graphlet degree distribution agreement

For each orbit j, one needs to measure the jth GDD, $dG_j(k)$, i.e., the distribution of the number of nodes in G "touching" the corresponding graphlet at orbit j k times. Clearly, the degree distribution is the 0th GDD.

$dG_j(k)$ is scaled as $S^j_G(k) = \frac{d^j_G(k)}{k}$ to decrease the contribution of larger degrees in a GDD and then normalized with respect to its total area $T^j_G = \sum_{k=1}^{\infty} S^j_G(k)$ giving the "normalized distribution"

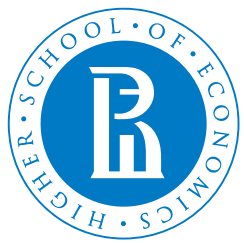
$$N^j_G(k) = \frac{S^j_G(k)}{T^j_G}$$

.

The jth GDD-agreement compares the jth GDDs of two networks. For two networks G and H and a particular orbit j, the "distance" $D_j(G, H)$ between their normalized jth GDDs is:

$$D^j(G, H) = \frac{1}{\sqrt{2}}(\sum_{k=1}^{\infty}[N^j_G(k) - N^j_H(k)]^2)^{\frac{1}{2}}$$

.



Graphlet degree distribution agreement

The distance is between 0 and 1, where 0 means that G and H have identical jth GDDs, and 1 means that their jth GDDs are far away. Next, $D_j(G,H)$ is reversed to obtain the jth GDD-agreement:

$$A^j(G,H) = 1 - D^j(G,H),$$

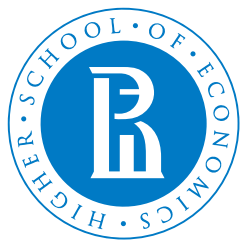
for $j \in \{0,1,\ldots,72\}$.
The total GDD-agreement between two networks G and H is the arithmetic or the geometric average of the jth GDD-agreements over all j, i.e.,

$$A_{arith}(G,H) = \frac{1}{73}\sum_{j=0}^{72} A^j(G,H),$$

and

$$A_{geo}(G,H) = \left(\prod_{j=0}^{72} A^j(G,H)\right)^{\frac{1}{73}},$$

respectively. GDD-agreement is scaled to always be between 0 and 1, where 1 means that two networks are identical with respect to this property.



Graphlet degree vectors (signatures) and signature similarities

The signature similarity is computed as follows. For a node u in graph G, u_i denotes the ith coordinate of its signature vector, i.e., u_i is the number of times node u is touched by an orbit i in G. The distance $D_i(u, v)$ between the ith orbits of nodes u and v is defined as:

$$D_i(u, v) = w_i \times \frac{|log(u_i + 1) - log(v_i + 1)|}{log(max\{u_i, v_i\} + 2)},$$

where wi is the weight of orbit i that accounts for dependencies between orbits. The total distance $D(u, v)$ between nodes u and v is defined as:

$$D(u, v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i}.$$

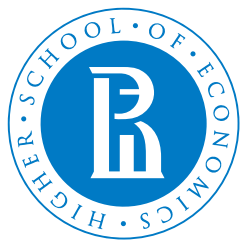
The distance $D(u, v)$ is in $[0, 1)$, where distance 0 means that signatures of nodes u and v are identical. Finally, the signature similarity, $S(u, v)$, between nodes u and v is:

$$S(u, v) = 1 - D(u, v).$$

Clearly, a higher signature similarity between two nodes corresponds to a higher topological similarity between their extended neighborhoods (out to distance 4).

Graphlets

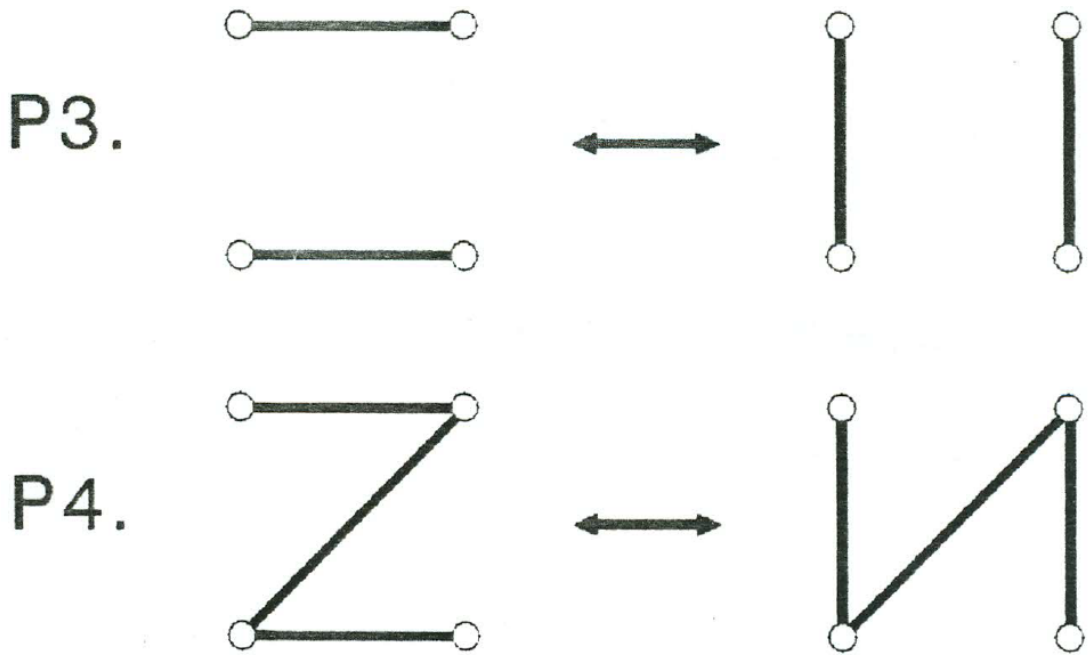
The notion of graphlets can be extended to directed networks but their number increases considerably paper1, paper2, paper3.
Orca / R



Inductive definitions

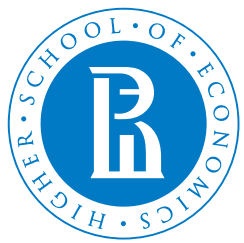
[5] [10]
Jason Vallet: Where Social Networks, Graph Rewriting and Visualisation Meet : Application to Network Generation and Information Diffusion page

Transformations



Preserve degrees in nodes.

A dissimilarity between two networks can be defined also as a shortest (weighted) transformations path leading from first to the second.



Resources I

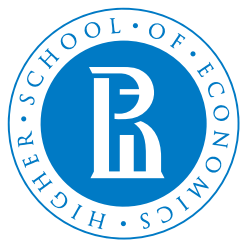
Batagelj, Vladimir : Similarity measures between structured objects. A. Graovac (ed.): Proceedings of MATH/CHEM/COMP 1988. Studies in Physical and Theoretical Chemistry, vol. 63. Amsterdam: Elsevier, 25-40 1989. paper

Batagelj, V, Bren, M: Comparing similarity measures. J Classif 12 (1): 73-90 1995 paper

Batagelj, V: Dissimilarities between structured objects – fragments. 1991. preprint

Batagelj, V., Mrvar, A. (2001). A subquadratic triad census algorithm for large sparse networks with small maximum degree. Social Networks, 23, 237-243.

Batagelj, V: Inductive classes of cubic graphs. Proceedings of the 6th Hungarian Colloquium on Combinatorics: Finite and infinite sets, Eger, Hungary, p. 89-101. paper



Resources II

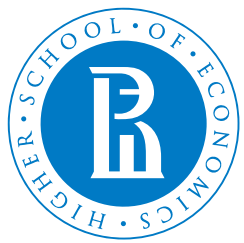
Bressan, Marco, Chierichetti, Flavio, Kumar, Ravi, Leucci, Stefano, Panconesi, Alessandro: Counting Graphlets: Space vs Time. WSDM '17 Proceedings of the Tenth ACM International Conference on Web Search and Data Mining Cambridge, United Kingdom — February 06 - 10, 2017, Pages 557-566

Deza, Michel Marie, Deza, Elena: Encyclopedia of Distances. Springer 2009.

Estrada, Ernesto, Knight, Philip A.: A First Course in Network Theory. Oxford UP, 2015.

Hočevar, Tomaž, Demšar, Janez: A combinatorial approach to graphlet counting. Bioinformatics, Volume 30, Issue 4, 15 February 2014, Pages 559–565, https://doi.org/10.1093/bioinformatics/btt717

Kejžar, N., Nikoloski, Z., Batagelj, V.: Probabilistic Inductive Classes of Graphs. Journal of Mathematical Sociology 32: 85-109, 2008. paper



Resources III

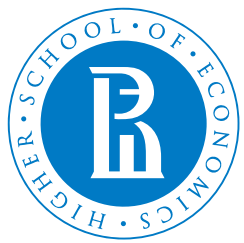
Keržič, D, Jerman Blažič, B, Batagelj, V: Comparison of three different approaches to the property prediction problem. J. Chem. Inf. Comput. Sci., 1994, 34 (2), pp 391–394. paper

Milo, R, Shen-Orr, S, Itzkovitz, S, Kashtan, N, Chklovskii, D, Alon, U: Network Motifs: Simple Building Blocks of Complex Networks. Science, 298, October 2002, p. 824-827.

Pržulj, Nataša: Biological network comparison using graphlet degree distribution. Bioinformatics, Volume 23, Issue 2, 15 January 2007, Pages e177–e183, https://doi.org/10.1093/bioinformatics/btl301

Wang, Cheng; Lizardo, Omar; Hachen, David; Strathman, Anthony; Toroczkai, Zoltan; Chawla, Nitesh V.: A dyadic reciprocity index for repeated interaction networks. Network Science, Vol 1, Issue 1 April 2013 , pp. 31-48.

Pajek: http://mrvar.fdv.uni-lj.si/pajek/



Resources IV

- Pajek history June 15, 1997: Reduction of flow graphs, searching fragments (in moleculae or graphs)
- Wikipedia: Motifs
- Wikipedia: /Graphlets