Drilling into Erasmus learning mobility flows between countries 2014-2024

Vladimir Batagelj

IMFM Ljubljana, UP IAM Koper, UL FMF Ljubljana

HiTEc & CoDES 2025

CUT, Limassol, Cyprus, 8-9 July 2025

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (July 3, 2025 at 16 : 55): slides PDF
https://github.com/bavla/wNets/

Erasmus

Erasmus+ is a European Union (EU) program designed to support education, training, youth, and sport across Europe. Established in 1987, it aims to provide opportunities for individuals to study, train, gain work experience, and volunteer abroad, while also fostering cooperation and innovation in these fields.

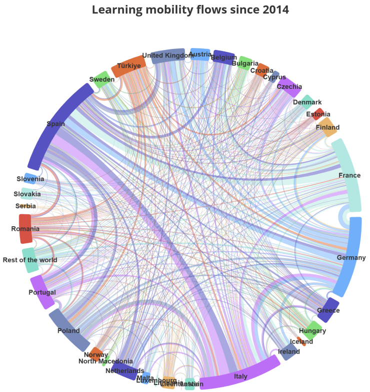
Key features of Erasmus+ are (1) mobility opportunities, (2) cooperation projects, (3) policy development, and (4) sport initiatives.

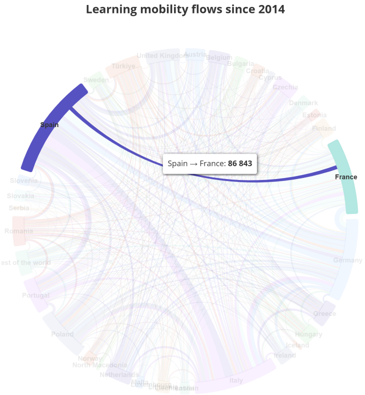
Erasmus+ is funded by the EU, with a budget of over €26 billion for the 2021-2027 period, making it one of the largest programs of its kind. It is open to EU member states, as well as non-EU countries associated with the program. Millions of individuals and thousands of organizations participate annually.

At the bottom of the Erasmus+ page Data visualization on learning mobility projects, the "Learning mobility flows since 2014" chart can be found.

Erasmus

interactive chart

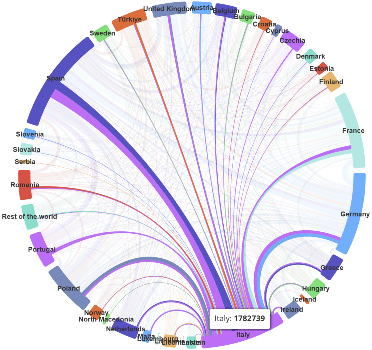


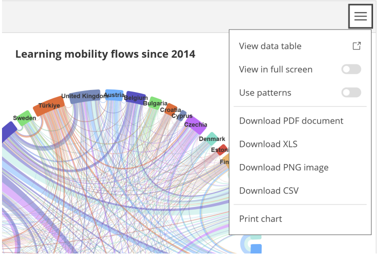


The interactive chart shows mobility flows between countries since 2014. The colors are related to the sending country.

Erasmus

. . . interactive chart





For example, moving the mouse over Italy will highlight all its in/outbound flows and the total count of participants. The same can be done at the flow level.

The interactive chart provides an option to download the network data.

Erasmus

creating network

I saved the network data on the file `Learning-mobility-flows-since-2014.csv`. The dataset contains the following countries:

Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands, North Macedonia, Norway, Poland, Portugal, Rest of the world, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Türkiye, United Kingdom.

I used the Deepseek to obtain the corresponding ISO 3166-1 alpha-2 country codes (`ISO2.csv`) and the total population estimate for each country (`pop.csv`).

I combined the collected data into an igraph network and saved it as `ErasmusFlows.rds`. The files are available at GitHub/Vlado.

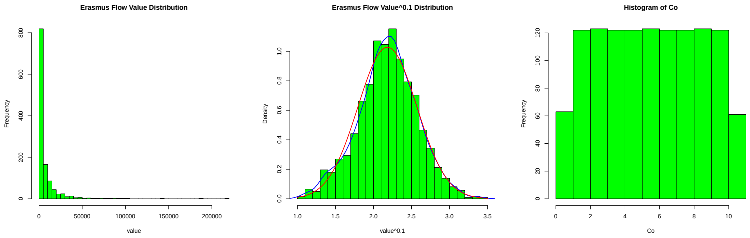
Network visualization

comments

- Larger, $n > 20$, dense graphs can't be presented readably with a graphical layout. For the Erasmus network, the number of nodes is $n = 35$, and the density $\gamma = 0.9984$. For dense graphs of moderate size (up to some hundreds of nodes) a better option is the matrix representation.
- What about weights? They can be represented by link thickness or level of grey of matrix cells. The problem is a very large range and the distribution of weights – most weights give almost white cells. For Erasmus $w_{\text{min}} = 1$ and $w_{\text{max}} = 217003$. Monotonic transformations such as $w' = a \cdot w$, $a > 0$ or $w' = \sqrt{w}$ or $w' = \log(w)$, etc. In our case, we used $w' = w^{0.1}$.

Erasmus network

. . . monotonic transformations and weight distributions



Monotonic (increasing) transformation $f : \mathbb{R} \to \mathbb{R}$

$$x < y \Rightarrow f(x) \leq f(y)$$

They preserve the ordering of weights.
Let $w' = f \circ w$ then $w(x) < w(y) \Rightarrow w'(x) \leq w'(y)$.

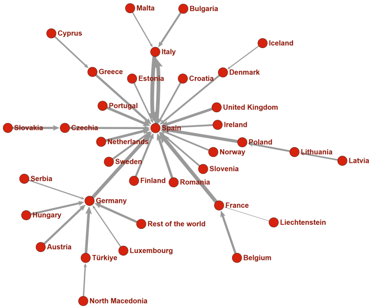
Skeletons

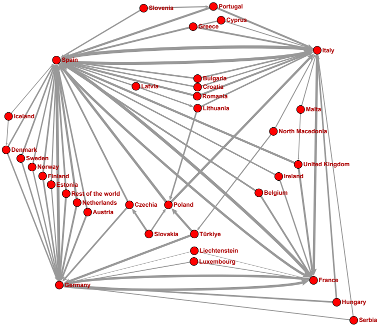
To get insight into the structure of a large (or/and) dense network, we can reduce it to its skeleton by removing less important links and/or nodes [Batagelj (2011)].

- Most often, the spanning tree, link cut, or node cut is used.
- In the closest k-neighbor skeleton for each node, only the largest k incident links are preserved. Invariant for monotonic transformations.
- The Pathfinder algorithm was proposed in the 1980s by Schvaneveldt [Schvaneveldt etal.(1988), Schvaneveldt(1990), Batagelj etal.(2014)]. It removes from the network with a **dissimilarity** weight all links that do not satisfy the triangle inequality – if a shorter path exists that connects the link's end nodes, then the link is removed.
- Cores are a very efficient tool to determine the most cohesive (active) subnetworks [Batagelj and Zaveršnik(2011)]. The subset of nodes $\mathbf{C} \subseteq \mathcal{V}$ induces a P_s core at level t if for all $v \in \mathbf{C}$ it holds $\mathrm{wdeg}_{\mathbf{C}}(v) \geq t$, and $\mathbf{C}$ is the maximum such subset.

1-neighbors and 2-neighbors

. . . first and second choice





Pathfinder

The Pathfinder procedure was proposed in the 1980s by Schvaneveldt [Schvaneveldt etal.(1988), Schvaneveldt(1990), Vavpetič etal.(2009)] for simplifying weighted networks, where the weight measures a **dissimilarity** between nodes.

It is based on Minkowski operation $a \boxed{r} b = \sqrt[r]{a^r + b^r}$. For $r = 1$, $r = 2$, and $r = \infty$ we get $a \boxed{1} b = a + b$, $a \boxed{2} b = \sqrt{a^2 + b^2}$, and $a \boxed{\infty} b = \max(a, b)$.

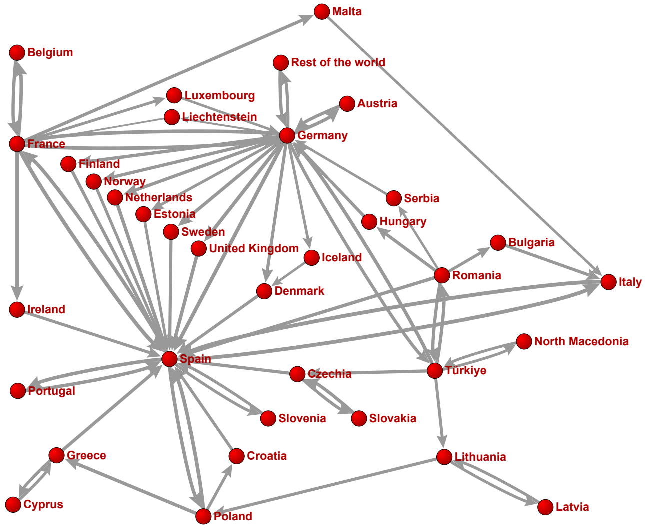
For a path $\pi = (v_1, v_2, \ldots, v_k)$ of length k we define its weight $w(\pi) = w(v_1, v_2) \boxed{r} w(v_2, v_3) \boxed{r} \ldots \boxed{r} w(v_{k-1}, v_k)$.

The Pathfinder procedure removes from a given network $\mathcal{N}$ every link (u, v) with its weight larger than the minimum weight of all u-v paths of length at most q. The resulting simplified network is denoted $\mathrm{PFnet}(\mathcal{N}, r, q)$.

Pathfinder

The Erasmus network weight w (number of visits) is a similarity measure. The Pathfinder procedure requires a dissimilarity measure d. A similarity w can be converted into a dissimilarity d in different ways. For example, $d_1 = w_{\max} - w$ or $d_2 = w_{\max}/w$. We will use the second option.

Erasmus network Pathfinder skeleton

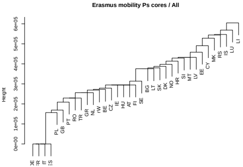


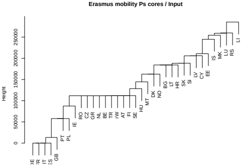
Observations

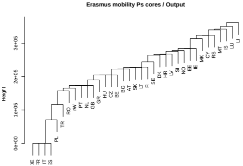
1. The 1-neighbors skeleton highlights Spain as the main attractor in the network.
2. The 2-neighbors skeleton shows the dominant role of Spain, Germany, France, and Italy.
3. The Pathfinder skeleton exposes Spain and Germany as central countries, and on the secondary level, France, Romania, Turkiye, Poland, and Italy.
4. These observations are confirmed by the Ps cores approach.

P_s-cores

Rank	All Id	All Value	Input Id	Input Value	Output Id	Output Value
1	DE	609063	DE	287693	DE	364594
2	FR	609063	FR	287693	FR	364594
3	IT	609063	IT	287693	IT	364594
4	ES	609063	ES	287693	ES	364594
5	PL	452314	GB	274340	PL	294156
6	GB	439822	PT	229822	TR	248328
7	PT	400014	PL	229822	RO	207249
8	RO	379701	IE	200266	rW	198970
9	TR	379701	RO	176038	PT	191225
10	GR	353090	CZ	176038	NL	191225
11	NL	353090	GR	176038	GB	191225
12	rW	339887	NL	176038	GR	174407
13	BE	336319	BE	176038	HU	159516
14	CZ	330134	TR	176038	CZ	159516
15	IE	314423	rW	175804	BE	159516
16	HU	314423	AT	175804	BG	141731
17	AT	314423	FI	175804	AT	141526
18	FI	314423	SE	175804	SK	136878
19	SE	295197	HU	159244	LT	136878
20	BG	233448	MT	143246	FI	136050
21	LT	233448	DK	125031	SE	120105
22	SK	229052	NO	125031	DK	100006
23	DK	221538	BG	103421	HR	98028
24	NO	211331	LT	103421	LV	96748
25	HR	195283	HR	103421	SI	88877
26	SI	179996	SK	99455	NO	86535
27	MT	176232	SI	99187	EE	80157
28	LV	176232	LV	81938	IE	80157
29	EE	150575	CY	81600	MK	50478
30	CY	118367	EE	76830	CY	40446
31	MK	80685	IS	42888	RS	40232
32	RS	64736	MK	33208	MT	24158
33	IS	62144	LU	28600	IS	21770
34	LU	40258	RS	27942	LU	13761
35	LI	4358	LI	2216	LI	2412







Matrix representation of Erasmus network

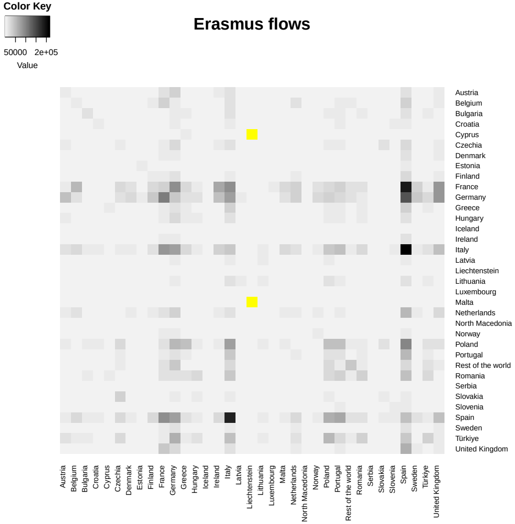
For dense graphs of moderate size (up to some hundreds of nodes) a better option is the matrix representation. Using the "right" ordering of its rows/columns we can obtain a picture in which the network structure is reflected as blocks in the matrix.

Such a "right" ordering of rows/cols in the matrix representation can be obtained by network clustering [Batagelj etal.(2014)]. Additional reordering of subtrees can be made manually using R or Pajek by reordering nodes in the hierarchy.

For visualization of matrices, we will use the function `heatmap.2` from the package `gplots`.

The weights can be represented by the level of grey or the color of matrix cells. The problem of a very large range and the distribution of weights remains – most weights give almost white cells. Again, we use monotonic transformations.

Matrix representation of Erasmus network



Matrix based (dis)similarities

For clustering units (nodes) we need a dissimilarity matrix D. In a square weight matrix, its weights can be sometimes considered (or transformed into) a dissimilarity.

$$D[u,v] = f(w[u,v], w[v,u]), \quad f(x,y) = f(y,x)$$

Often we use rows (and columns) as node descriptions and apply a selected dissimilarity on them

$$D[u,v] = d(w[u,.], w[v,.])$$

Typical dissimilarities are the *Euclidean distance*

$$d_e(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^2}$$

and the *Salton* or *cosine index*

$$S(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x} \bullet \mathbf{y}}{\sqrt{\mathbf{x}^2 \cdot \mathbf{y}^2}}, \ d_s(\mathbf{x},\mathbf{y}) = 1 - S(\mathbf{x},\mathbf{y}) \ or \ d_a(\mathbf{x},\mathbf{y}) = \frac{\arccos S(\mathbf{x},\mathbf{y})}{\pi}$$

where $\mathbf{x} \bullet \mathbf{y} = \sum_i x_i \cdot y_i$ and $\mathbf{x}^2 = \mathbf{x} \bullet \mathbf{x}$.

Corrected (dis)similarities

$$w[u,\cdot] = [w[u,1],\ldots,w[u,i],\ldots,w[u,u],\ldots,w[u,v],\ldots,w[u,k]]$$

$$w[v,\cdot] = [w[v,1],\ldots,w[v,i],\ldots,w[v,u],\ldots,w[v,v],\ldots,w[v,k]]$$

In traditional (dis)similarities, comparing $w[u,i]$ and $w[v,i]$ we are comparing how u relates to i with how v relates to i. The problem arises for $i = u$ and $i = v$. We would need to compare $w[u,u]$ with $w[v,v]$ and $w[u,v]$ with $w[v,u]$. This leads to *corrected* (dis)similarities.

Corrected Euclidean distance and Salton index

Corrected Euclidean distance

$$d'_e(u,v) = \sqrt{(w[u,v] - w[v,u])^2 + (w[u,u] - w[v,v])^2 + \sum_{t \notin \{u,v\}} (w[u,t] - w[v,t])^2}$$

Corrected Salton index of the link $(u,v) \in \mathcal{L}$

$$S'(u,v) = \frac{w[u,.] \bullet w[v,.] + (w[u,u] - w[u,v]) \cdot (w[v,v] - w[v,u])}{\sqrt{w[u,.]^2 \cdot w[v,.]^2}}$$

It has the following properties

1. $S'(u,v) \in [-1,1]$
2. $S'(u,v) = S'(v,u)$
3. $S'(u,u) = 1$
4. $w : L \to \mathbb{R}_0^+ \Rightarrow S'(u,v) \in [0,1]$
5. $S'(\alpha u, \beta v) = S'(u,v), \quad \alpha, \beta > 0$
6. $S'(\alpha u, u) = 1, \quad \alpha > 0$

Normalizations

activity or Balassa index

In networks with weights with a large range usually a few strong nodes prevail. To diminish or neutralize the influence of size on results different normalizations were proposed and used [Batagelj and Mrvar(2003), Matveeva etal.(2023)].

Let $T = \sum_{e \in \mathcal{L}} w(e)$ and for $(u, v) \in \mathcal{L}$ (Balassa index)

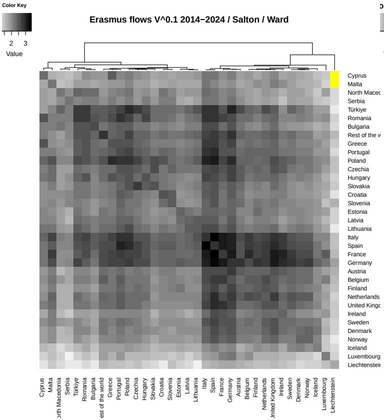
$$A(u, v) = \frac{w[u, v] \cdot T}{\text{woutdeg}(u) \cdot \text{windeg}(v)}$$

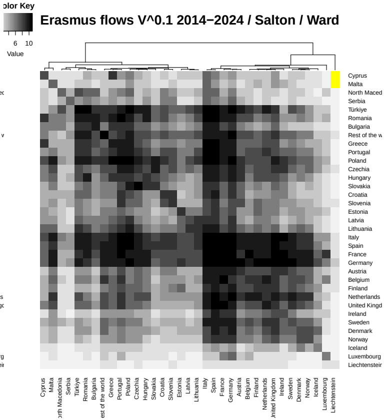
then the *activity normalization* w'

$$w'(u, v) = \log_2 A(u, v)$$

Erasmus mobilty flow matrix

Salton clustering



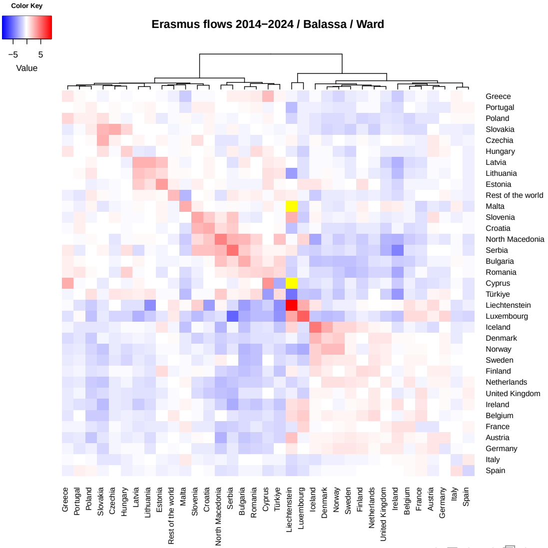


Observations

1. The "cross" formed by C_1 = (Italy, Spain, France, Germany) – strong activity with almost all countries in both directions.
2. Intense diagonal "squares" – clusters: C_2 = (Türkiye, Romania, Bulgaria, Rest of the world, Greece, Portugal, Poland, Czechia, Hungary), C_3 = (Poland, Czechia, Hungary, Slovakia), C_4 = (Croatia, Slovenia), C_5 = (Estonia, Latvia, Lithuania), C_6 = (Austria, Belgium, Finland, Netherlands, United Kingdom, Ireland, Sweden, Denmark, Norway), $C_1 \cup C_6$, C_7 = (Sweden, Denmark, Norway, Iceland)
3. Out-diagonal "rectangles": Luxembourg $\times$ (France, Germany), Greece $\times$ Cyprus, (Croatia, Slovenia) $\times$ (North Macedonia, Serbia), etc.
4. In the cross, $C_4 \cup C_5 \cup$ Slovakia less often select France, etc.

Erasmus mobility flow matrix

Balassa clustering



Observations

1. Three main clusters B_l = (Greece : Türkiye) – less developed, B_h = (Iceland : Spain) – high developed. and B_L =(Liechtenstein, Luxembourg). Most cells inside squares are red and out-diagonal rectangles are mostly blue – exchange between countries from the same cluster is above expected, and below expected between different clusters.

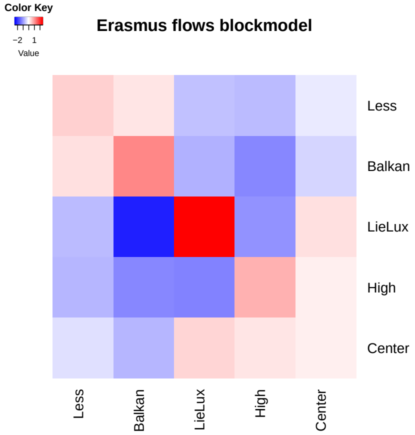
2. Red diagonal "squares" – clusters: B_1 = (Slovakia, Czechia, Hungary), B_2 =(Latvia, Lithuania, Estonia), B_3 = (Slovenia, Croatia, North Macedonia, Serbia), B_4 = (North Macedonia, Serbia, Bolgaria, Romania), B_L. The exchange between Cyprus and Türkiye is bellow expected. In the main cluster B_h we can identify a subcluster B_5 = (Iceland, Denmark, Norway, Sweden, Finland, Netherlands, United Kingdom). Within the clusters B_1, B_2, B_3, B_4, B_L, and (Iceland, Denmark, Norway) visits are much above expected.

Observations

3. Countries from the cluster B_I are selecting Malta bellow expected. The exchange between B_5 and B_L is bellow expected.
4. Exchange between Cyprus and Greece is above expected.
5. Exchange of Italy, Spain and Estonia with other countries is mostly close to as expected.

Erasmus mobilty flow blockmodel

Balassa clustering



1. **Less:** Greece, Portugal, Poland, Slovakia, Czechia, Hungary, Latvia, Lithuania, Estonia, Rest of the world, Malta,
2. **Balkan:** Slovenia, Croatia, North Macedonia, Serbia, Bulgaria, Romania, Cyprus, Türkiye,
3. **LieLux:** Liechtenstein, Luxembourg,
4. **High:** Iceland, Denmark, Norway, Sweden, Finland, Netherlands, United Kingdom,
5. **Center:** Ireland, Belgium, France, Austria, Germany, Italy, Spain.

Conclusions

1. Why is Spain the most attractive country?
2. How can the blue between less and high developed countries be reduced?
3. This is exploratory network analysis. Collect and use additional data (neighbors relation, population size, GDP, etc.).
4. Temporal version of the network.

Acknowledgments

The computational work reported in this presentation was performed using R. The code and data are available at GitHub/Vlado.

This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research project J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc).

References I

- Batagelj, V, Zaveršnik, M: Fast algorithms for determining (generalized) core groups in social networks. Advances in Data Analysis and Classification, 2011. Volume 5, Number 2, 129-145.
- Batagelj, V, Doreian, P, Ferligoj, A, Kejžar, N (2014). Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley. ISBN: 978-0-470-71452-2
- Batagelj, V (2011). Large-Scale Network Analysis. in John Scott, Peter J. Carrington eds. The SAGE Handbook of Social Network Analysis SAGE Publications.
- Batagelj, V, & Mrvar, A (2003). Density based approaches to network analysis; Analysis of Reuters terror news network. Workshop on link analysis for detecting complex behavior (LinkKDD2003). Retrieved August 27, 2003, from LinkKDD2003.
- Matveeva, N, Batagelj, V, Ferligoj, A (2023). Scientific collaboration of post-Soviet countries: the effects of different network normalizations Scientometrics, Volume 128, issue 8, Pages: 4219 – 4242
- Schvaneveldt, RW (Editor) (1990). Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex. book.

References II

- Schvaneveldt, RW, Dearholt, DW, Durso, FT (1988). Graph theoretic foundations of Pathfinder networks. Computers and Mathematics with Applications, 15, 337-345.
- Vavpetič, A, Batagelj, V, Podpečan, V (2009). "An implementation of the Pathfinder algorithm for sparse networks and its application on text network." In 12th International Multiconference Information Society, vol. A, pp. 236-239. paper.