



Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

Networks and symbolic data analysis

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE ANR Moscow

Final COSTNET conference

Zoom, 24-25. September 2020



Outline

Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

- 1 Symbolic data analysis
- 2 Semirings and networks
- 3 Example: Social Networks
- 4 Example: Terror news
- 5 References

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

Last version of slides (September 24, 2020, 12:59): [symbnet.pdf](#)



Symbolic data analysis

Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

An approach to deal with big data is to aggregate the data into a smaller, manageable data set that can be analyzed using standard data analysis methods.

In *symbolic data analysis* a variable is not aggregated into a single number (mean value), but its values are summarized using complex data structures (for example a histogram or a temporal quantity) preserving more information. Symbolic data analysis provides methods for analysis of so obtained complex data frames.

Such complex data frames can be obtained also in other ways.

- time series (John Graunt, 1662 [[11](#)])
- symbolic data analysis (Edwin Diday, 1987 [[9](#), [6](#)])
- compositions (John Aitchison, 1982 [[1](#), [2](#)])
- functional data analysis (James Ramsay, 1982 [[14](#), [15](#)])
- object (oriented) data (H. Wang and Steve Marron, 2007 [[17](#), [12](#)])

In complex data analysis the measured values over a selected group A are aggregated into a complex object $\Sigma(A)$ and not into a single value. An interesting question is, which complex data types are *compatible with merging* of disjoint sets of units

$$\Sigma(A \cup B) = F(\Sigma(A), \Sigma(B)), \quad \text{for } A \cap B = \emptyset.$$

We assume $A \cap B = \emptyset$

- 1 $\Sigma(A) = |A| = n_A, \quad \Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$
- 2 $\Sigma(A) = \min_{X \in A} v(X), \quad \Sigma(A \cup B) = \min(\Sigma(A), \Sigma(B))$
- 3 $\Sigma(A) = \max_{X \in A} v(X), \quad \Sigma(A \cup B) = \max(\Sigma(A), \Sigma(B))$
- 4 $\Sigma(A) = (\text{First}(A), \text{Second}(A)),$
 $\Sigma(A \cup B) = (\text{First}(L), \text{Second}(L)),$ where
 $L = \{\text{First}(A), \text{Second}(A), \text{First}(B), \text{Second}(B)\}$
- 5 $\Sigma(A) = (n_A, \mu_A), \quad \Sigma(A \cup B) = (n_A + n_B, \frac{n_A \mu_A + n_B \mu_B}{n_A + n_B})$
- 6 $\Sigma(A) = \sum_{X \in A} v(X), \quad \Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$

An aggregation of numerical variable $v(X)$ realized as a vector \mathbf{x} is represented with a triple (n_x, μ_x, σ_x) . It is an exactly mergeable summary. Holds also for higher order moments.

Counting number of members/values from C in A :
 $n(A; C) = |A \cap C|$ is an exactly mergeable summary

$$n(A \cup B; C) = n(A; C) + n(B; C)$$

Let Σ_1 and Σ_2 be exactly mergeable summaries. Then also

$$\Sigma(A) = \Sigma_1 \oplus \Sigma_2(A) = (\Sigma_1(A), \Sigma_2(A))$$

is an exactly mergeable summary.

Therefore, since set membership counts are exactly mergeable, the *barcharts* $C = \{X : v(X) = c\}$ and *histograms* $C = \{X : v(X) \in [a, b]\}$ are exactly mergeable summaries.

Intervals $\Sigma(A) = [\min_{X \in A} v(X), \max_{X \in A} v(X)]$ are exactly mergeable summaries.



Symbolic networks

Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

A *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ is a *symbolic network* if some property of nodes from \mathcal{P} or some weight on links from \mathcal{W} has symbolic values.

Let $a, b, c \in A$. The set A with binary operations addition \oplus and multiplication \odot , neutral element 0 and unit 1 , denoted with $A(\oplus, \odot, 0, 1)$, is a *semiring*, when the following conditions hold:

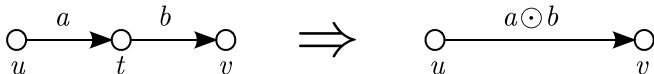
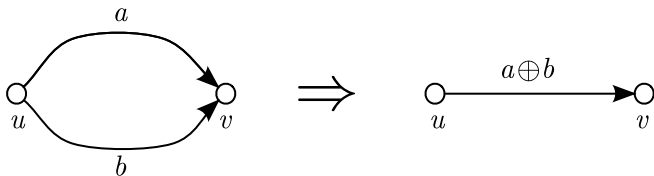
- the set A is a *abelian monoid* for the addition \oplus with a neutral element 0 (the addition is commutative, associative and $a \oplus 0 = a$ for all $a \in A$);
- the set A is a *monoid* for the multiplication \odot with the unit 1 (the multiplication is associative and $a \odot 1 = 1 \odot a = a$ for all $a \in A$);
- the addition *distributes* over the multiplication

$$a \odot (b \oplus c) = (a \odot b) \oplus (a \odot c) \quad \text{and} \\ (a \oplus b) \odot c = (a \odot c) \oplus (b \odot c);$$

- the element 0 is an absorbing element or *zero* for the multiplication: $a \odot 0 = 0 \odot a = 0$ for all $a \in A$.

In all cases we assume precedence of the multiplication over the addition. The last point in the definition of semirings is omitted by some authors.

Semirings are the right structure for extending the weights from links to nodes, walks (paths) and sets of walks.



- 1 Combinatorial: $(\mathbb{N}, +, \cdot, 0, 1)$ or $(\mathbb{R}_0^+, +, \cdot, 0, 1)$
- 2 Shortest paths: $(\mathbb{R}_0^+, \min, +, \infty, 0)$
- 3 Interval 1 [13]: $[a, A], [b, B] \subset \mathbb{R}_0^+$
 $[a, A] \oplus [b, B] = [a + b, A + B]$ and
 $[a, A] \odot [b, B] = [a \cdot b, A \cdot B]$
- 4 Interval 2: $[a, A], [b, B] \subset \mathbb{R}$
 $[a, A] \oplus [b, B] = [\min(a, b), \max(A, B)]$ and
 $[a, A] \odot [b, B] = [a + b, A + B]$
- 5 Interval 3:

$$[a, A] \oplus [b, B] = \begin{cases} [a, A] & A < B \\ [b, B] & B < A \\ [\min(a, b), A] & A = B \end{cases}$$

Let the set of bins $\mathbf{B} = \{B_1, B_2, \dots, B_k\}$ be a partition of the set B .

A *histogram* $h : \mathbf{B} \rightarrow \mathbb{N}$

$$h_i = h(B_i) = |\{X : v(X) \in B_i\}|$$

$$h \oplus g = h + g \quad (h \oplus g)(i) = h(i) + g(i)$$

$$h \odot g = h * g \quad \text{convolution [10, 7]} \quad (h * g)(i) = \sum_{p \circ q = i} h(p) \cdot g(q)$$

○ is a semigroup operation on bins

A temporal quantity (TQ) a is a function $a : \mathcal{T} \rightarrow A \cup \{\mathbb{K}\}$ where \mathbb{K} denotes the value *undefined*. The *activity time set* T_a of a consists of instants $t \in \mathcal{T}$ in which a is defined $T_a = \{t \in \mathcal{T} : a(t) \in A\}$.

We can extend both operations to the set $A_{\mathbb{K}} = A \cup \{\mathbb{K}\}$ by requiring that for all $a \in A_{\mathbb{K}}$ it holds $a + \mathbb{K} = \mathbb{K} + a = a$ and $a \cdot \mathbb{K} = \mathbb{K} \cdot a = \mathbb{K}$. The structure $(A_{\mathbb{K}}, +, \cdot, \mathbb{K}, 1)$ is also a semiring.

Let $A_{\mathbb{K}}(\mathcal{T})$ denote the set of all TQs over $A_{\mathbb{K}}$ in time \mathcal{T} . To extend the operations to networks and their matrices we first define the *sum* (parallel links) $a + b$ as

$$(a + b)(t) = a(t) + b(t) \quad \text{and} \quad T_{a+b} = T_a \cup T_b.$$

The *product* (sequential links) $a \cdot b$ is defined as

$$(a \cdot b)(t) = a(t) \cdot b(t) \quad \text{and} \quad T_{a \cdot b} = T_a \cap T_b.$$



Temporal quantities

Symbolic networks

V. Batagelj

Symbolic data analysis

Semirings and networks

Example:
Social Networks

Example:
Terror news

References

Let us define TQs $\mathbf{0}$ and $\mathbf{1}$ with requirements $\mathbf{0}(t) = \mathbb{K}$ and $\mathbf{1}(t) = 1$ for all $t \in \mathcal{T}$. Again, the structure $(A_{\mathbb{K}}(\mathcal{T}), +, \cdot, \mathbf{0}, \mathbf{1})$ is a semiring.

To produce a software support for computation with TQs we limit it to TQs that can be described as a sequence of disjoint time intervals with a constant value

$$a = [(s_i, f_i, v_i)]_{i \in 1..k}$$

where s_i is the starting time and f_i the finishing time of the i -th time interval $[s_i, f_i)$, $s_i < f_i$ and $f_i \leq s_{i+1}$, and v_i is the value of a on this interval (over combinatorial semiring). Outside the intervals the value of TQ a is undefined, \mathbb{K} . Therefore

$$T_a = \bigcup_{i \in 1..k} [s_i, f_i).$$

Let the binary *affiliation* matrix $\mathbf{A} = [a_{ep}]$ describe a two-mode network on the set of events E and the set of participants P :

$$a_{ep} = \begin{cases} 1 & p \text{ participated at the event } e \\ 0 & \text{otherwise} \end{cases}$$

The function $d : E \rightarrow \mathcal{T}$ assigns to each event e the date $d(e)$ when it happened. Assume $\mathcal{T} = [first, last] \subset \mathbb{N}$. Using these data we can construct two temporal affiliation matrices:

- **instantaneous** $\mathbf{A}_i = [a_{i_{ep}}]$, where

$$a_{i_{ep}} = \begin{cases} [(d(e), d(e) + 1, 1)] & a_{ep} = 1 \\ [] & \text{otherwise} \end{cases}$$

- **cumulative** $\mathbf{A}_c = [a_{c_{ep}}]$, where

$$a_{c_{ep}} = \begin{cases} [(d(e), last + 1, 1)] & a_{ep} = 1 \\ [] & \text{otherwise} \end{cases}$$

Let \mathbf{N} be a temporal network on $E \times P$. On it we can define some interesting temporal quantities [4] such as *in-sum*:

$$iS(\mathbf{N}, p) = \sum_{e \in E} n_{ep}$$

and *out-sum*:

$$oS(\mathbf{N}, e) = \sum_{p \in P} n_{ep}$$

For $\mathbf{N} \equiv \mathbf{W}\mathbf{A}\mathbf{i}$ (W – set of works; A – set of authors) we get the *productivity of an author* a :

$pr(a) = iS(\mathbf{W}\mathbf{A}\mathbf{i}, a)$ = number of publications of the author a by year

and for $\mathbf{N} \equiv \mathbf{W}\mathbf{A}\mathbf{c}$ we get the *cumulative productivity of an author* a :

$cpr(a) = iS(\mathbf{W}\mathbf{A}\mathbf{c}, a)$ = cumulative number of publications of the author a by year.

The productivity of an author can be extended to the *productivity of a group of authors* C

$$pr(C) = \sum_{a \in C} pr(a) = \sum_{a \in C} iS(\mathbf{WA}i, a)$$

There is a problem with the productivity of a group. In the case when two authors from a group co-authored the same paper it is counted twice. To account for a “real” contribution of each author the fractional approach is used. It is based on normalized networks (matrices) – in the case of co-authorship on $n(\mathbf{WA}) = \mathbf{WAN} = [wan_{wa}]$

$$wan_{wa} = \frac{wa_{wa}}{\max(1, \text{outdeg}_{\mathbf{WA}}(w))}.$$

This leads to the *fractional productivity of an author* a :

$fpr(a) = iS(\mathbf{WAn}i, a)$ = fractional contribution of publications of the author a by year

Derived networks – citations between journals

The *network multiplication* $N_C = N_A * N_B$ over a selected semiring is defined in a standard way

$$c[u, v] = \bigoplus_{z \in N(u) \cap N^-(v)} a[u, z] \odot b[z, v]$$

It is about traveling (sets of walks) in the network. It is also used to get the derived networks from basic networks. For example:

Based on temporal networks **WJins**, **WJcum**, and **Citelns**, we constructed two types of temporal networks of *citations between journals* **JCJ** and **JCJn**.

$$\mathbf{JCJ} = \mathbf{WJins}^T * \mathbf{Citeln} * \mathbf{WJcum}$$

$$\mathbf{JCJn} = \mathbf{WJins}^T * n(\mathbf{Citeln}) * \mathbf{WJcum}$$

The first network counts the number of citations between journals, and the second contains the corresponding fractional values.



Social networks bibliography

self-citation, SocNet → AJSoC, in, out

Symbolic networks

V. Batagelj

Symbolic data analysis

Semirings and networks

Example: Social Networks

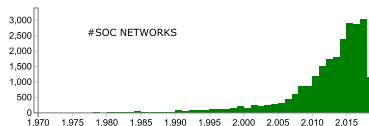
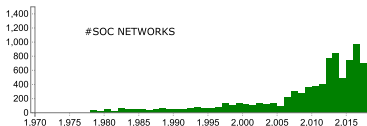
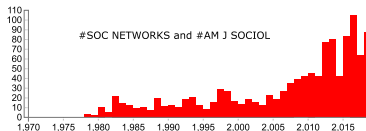
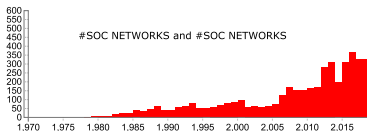
Example: Terror news

References

70792 works, 93011 authors, 8943 journals, 32409 keywords / complete description

$$\text{self-cite}(\text{SocNet}) = \mathbf{JCJ}[\text{SocNet}, \text{SocNet}]$$

$$\text{SocNet} \rightarrow \text{AJSoC} = \mathbf{JCJ}[\text{SocNet}, \text{AJSoC}]$$



$$\text{citing}(\text{SocNet}) = oS(\mathbf{JCJ}, \text{SocNet})$$

$$\text{cited}(\text{SocNet}) = iS(\mathbf{JCJ}, \text{SocNet})$$



For a unit X_i , each variable V_j is described with a size h_{ij} and a TQ \mathbf{x}_{ij}

$$X_{ij} = (h_{ij}, \mathbf{x}_{ij})$$

In our algorithms we use *normalized* values of temporal variables $V' = (h, \mathbf{p})$ where

$$\mathbf{p} = [(s_r, f_r, p_r) : r = 1, 2, \dots, k] \quad \text{and} \quad p_r = \frac{v_r}{h}$$

In the case, when $h = \text{tot}(\mathbf{x}) = \sum v_r$, the normalized TQ \mathbf{p} is essentially a probability distribution.

For clustering TQs we implemented the leaders method and agglomerative hierarchical clustering method. Both methods are compatible – they are based on the same clustering error criterion function.



September 11th Reuters terror news network

Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

The Reuters terror news network **TN** was obtained from the CRA (Centering Resonance Analysis) networks produced by Steve Corman and Kevin Dooley at Arizona State University. The network is based on all the stories released during 66 consecutive days by the news agency Reuters concerning the September 11 attack on the U.S., beginning at 9:00 AM EST 9/11/01.

The nodes of this network are important words (terms). There is an edge between two words iff they appear in the same utterance [8]. The weight of an edge is its frequency. The network has $n = 13332$ nodes (different words in the news) and $m = 243447$ edges, 50859 with value larger than 1. There are no loops in the network.

To cluster all 13332 words (nodes) in Terror news described with $iS(\mathbf{TN}, u)$, $u \in V$ we first used the adapted leaders method searching for 100 clusters. After 50 steps we stopped the search. We continued with hierarchical clustering of the obtained leaders.



Hierarchical clustering

Symbolic networks

V. Batagelj

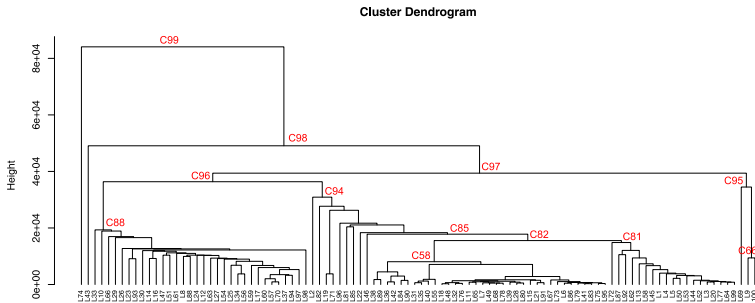
Symbolic data analysis

Semirings and networks

Example: Social Networks

Example: Terror news

References





Comparisons of leaders and cluster representatives L74:C98, C58:C81, C96:C95, C88:C94

Symbolic
networks

V. Batagelj

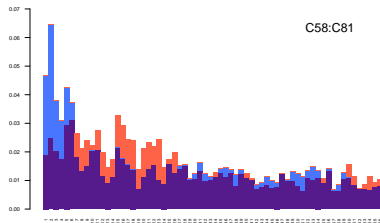
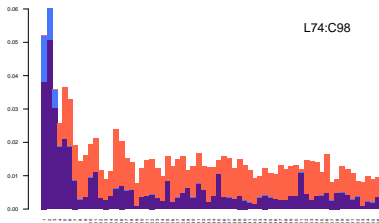
Symbolic data
analysis

Semirings and
networks

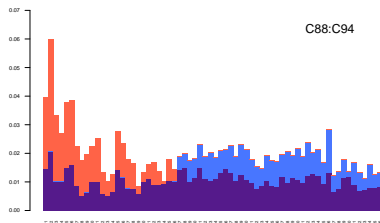
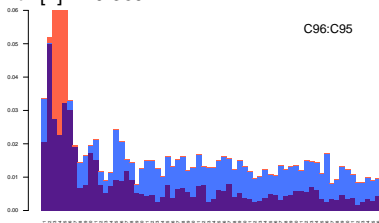
Example:
Social
Networks

Example:
Terror news

References










$$L74[2] = 0.5894$$





$$C95[3] = 0.1665, \quad C95[4] = 0.1570, \quad C95[5] = 0.2250$$


- NetsJSON – network description in JSON
<https://github.com/bavla/NetsJSON>
- TQ – basic support for temporal quantities in Python
<https://github.com/bavla/TQ>
- Nets - network analysis algorithms in Python
<https://github.com/bavla/Nets>
- Clustering of TQs in Python and R
<http://vladowiki.fmf.uni-lj.si/doku.php?id=vlado:work:alg:ldtq>

- 
 Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*. 44 (2): 139–177
- 
 Aitchison, J. (2011) [1986]. *The Statistical Analysis of Compositional Data*, Monographs on statistics and applied probability, Springer
- 
 Batagelj, V., Kejžar, N., and Korenjak-Černe, S. (2015). Clustering of Modal Valued Symbolic Data. *ArXiv e-prints*, [1507.06683](#).
- 
 Batagelj, V., Maltseva, D.(2020) Temporal bibliographic networks. *Journal of Informetrics*, Volume 14, Issue 1, 101006.
- 
 Batagelj, V., Praprotnik, S. (2016) An algebraic approach to temporal network analysis based on temporal quantities. *Social Network Analysis and Mining*, 6(1), 28.
- 
 Billard, L., Diday, E. (2012). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons.








- 

Horst Brunotte: Discrete Convolution Rings. Romanian Journal of Mathematics and Computer Science, 2013, Volume 3, Issue 2, p.155-159
- 

Corman, S.R., Kuhn, T., McPhee, R.D., Dooley, K.J. (2002) Studying complex discursive systems: Centering resonance analysis of communication. Human Communication Research, 28(2), 157-206.
- 

Diday E. (1987). The symbolic approach in clustering and related methods of Data Analysis. in Classification and Related Methods of Data Analysis, Proc. IFCS, Aachen, Germany. H. Bock ed. North-Holland.
- 

Dongol, B., Hayes, I.J., Struth, G. (2016). Convolution as a Unifying Concept: Applications in Separation Logic, Interval Calculi and Concurrency ACM Transactions on Computational Logic, February 2016, Article No.: 15

- 
 Graunt, J. (1662). Natural and Political Observations . . . Made upon the Bills of Mortality. London.
- 
 Marron, J.S., and Dryden, I.L. (2020). Object Oriented Data Analysis.
- 
 Moore, R.E., Kearfott, R.B., Cloud, M.J. (2009). Introduction to Interval Analysis. Philadelphia: SIAM.
- 
 Ramsay J.O. (1982). When the data are functions. *Psychometrika* 47:379–396.
- 
 Ramsay, J. O. and Silverman, B.W. (2005). Functional data analysis, 2nd ed., New York: Springer.
- 
 Rodriguez, O.R. (2018). *RSDA 2.0.5: R to Symbolic Data Analysis*.
<https://cran.r-project.org/web/packages/RSDA/>.
- 
 Wang, H. and Marron, J. S. (2007). Object oriented data analysis: sets of trees. *The Annals of Statistics* 35, 1849–1873



Acknowledgments

Symbolic
networks

V. Batagelj

Symbolic data
analysis

Semirings and
networks

Example:
Social
Networks

Example:
Terror news

References

This work was supported in part by the Slovenian Research Agency (research programs P1-0294 and research projects J5-2557, J1-9187 and J1-2481), project COSTNET (COST Action CA15109), and by Russian Academic Excellence Project '5-100'.