Semirings and temporal network analysis

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE ANR Moscow

8th European Congress of Mathematics

Portorož / Zoom, 20-26. June 2021

Outline

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

Last version of slides (June 22, 2021, 04 : 00): TempNA.pdf

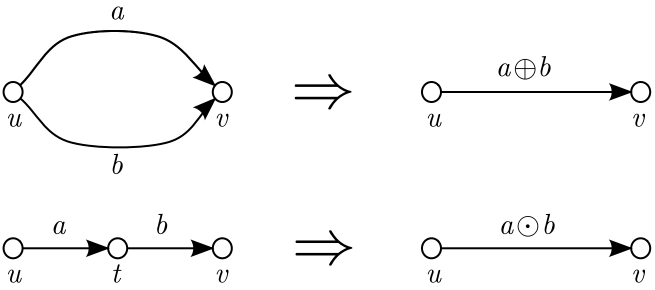
Networks

A *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of:

- a *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{A}$ is the set of arcs, $\mathcal{E}$ is the set of edges, and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$ is the set of links.
  $n = |\mathcal{V}|$, $m = |\mathcal{L}|$
- $\mathcal{P}$ *node value functions* / properties: $p : \mathcal{V} \to A$
- $\mathcal{W}$ *link value functions* / weights: $w : \mathcal{L} \to B$

Semirings

We would like to extend a weight $w \in \mathcal{W}$ to subsets of links, walks, and sets of walks. We need two operations on weights



Semirings are the right algebraic structure for this.

Semirings

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 $a \in A$.

Examples of semirings

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

Temporal quantities

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

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

Let $A_{\maltese}(\mathcal{T})$ denote the set of all TQs over $A_{\maltese}$ 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

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

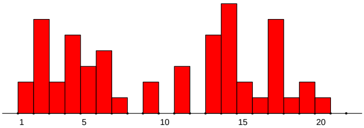
$$a = [(s_i, f_i, v_i)]_{i \in 1..k}$$

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 undedined, ⌘. Therefore

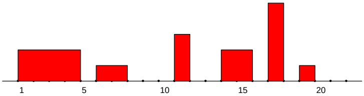
$$T_a = \bigcup_{i \in 1..k} [s_i, f_i).$$

Sum and product of temporal quantities

$a + b$:

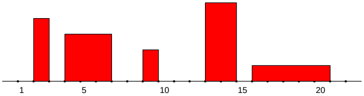


a :



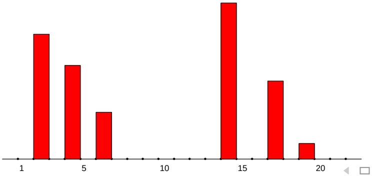
$$= [(1, 5, 2), (6, 8, 1), (11, 12, 3), (14, 16, 2), (17, 18, 5), (19, 20, 1)]$$

b :



$$= [(2, 3, 4), (4, 7, 3), (9, 10, 2), (13, 15, 5), (16, 21, 1)]$$

$a \cdot b$:



Instantaneous and cumulative temporal networks

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

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

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

- **instantaneous Ai** $= [ai_{ep}]$, where

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

- **cumulative Ac** $= [ac_{ep}]$, where

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

Node activity

Let **N** be a temporal network on $E \times P$. On it we can define some interesting temporal quantities [2] such as *in-sum*:

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

and *out-sum*:

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

n_{ep} is the temporal quantity assigned to the link (e, p).

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

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

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

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

Fractional activity

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

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

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 $n(\mathbf{WA}) = \mathbf{WAn} = [wan_{wa}]$

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

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

$fpr(a) = iS(\mathbf{WAni}, a) =$ 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 **CiteIns**, we constructed two types of temporal networks of *citations between journals* **JCJ** and **JCJn**.

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

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

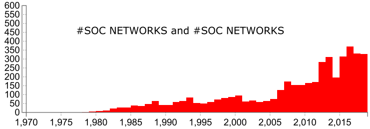
The first network counts the number of citations between journals, and the second contains the corresponding fractional values.

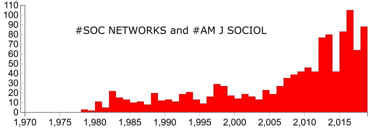
Social networks bibliography

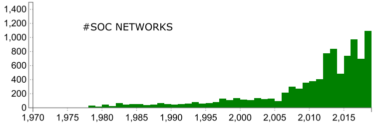
self-citation, SocNet $\to$ AJSoc, in, out

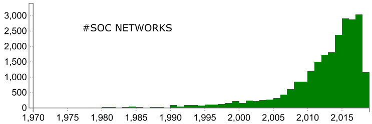
70792 works, 93011 authors, 8943 journals, 32409 keywords / complete description

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









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

Clustering of temporal quantities

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

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

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.

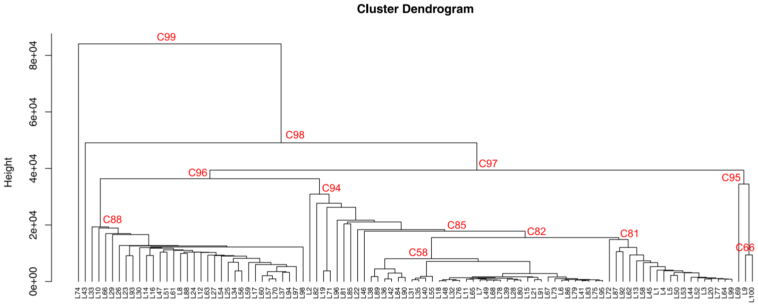
September 11th Reuters terror news network

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.

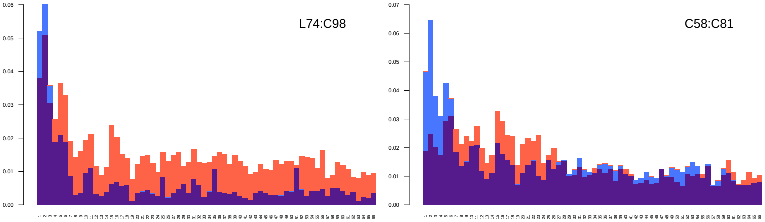
The nodes of this network are important words (terms). There is an edge between two words iff they appear in the same utterance [4]. The weight of an edge is its frequency. The network has $n = 13332$ nodes (different words in the news) and $m = 243447$ edges, 50859 with value larger than 1. There are 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.

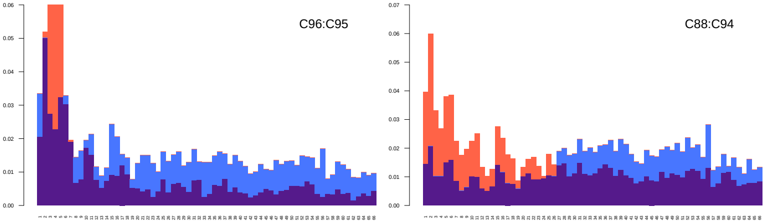
Hierarchical clustering



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



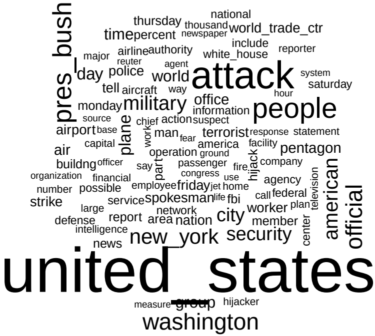
$L74[2] = 0.5894$



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

Word clouds for C58 and C81

$|C58| = 1396, \quad |C81| = 2226$



Software

- 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

References I

- Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2021). Clustering of Modal-Valued Symbolic Data. Advances in Data Analysis and Classification 15, pages 513–541
- 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.
- 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.
- Moore, R.E., Kearfott, R.B., Cloud, M.J. (2009). Introduction to Interval Analysis. Philadelphia: SIAM.

Acknowledgments

This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J1-9187 and J5-2557), and prepared within the framework of the HSE University Basic Research Program.