Signed two-mode networks

Salton's cosine

Jaccard's similarity

Remarks

References

Projections of signed two-mode networks

Vladimir Batagelj

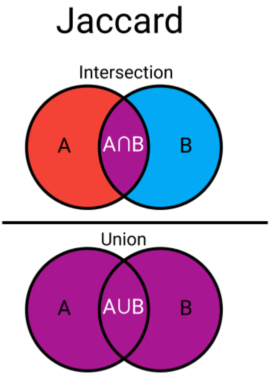
IMFM Ljubljana, IAM UP Koper, and NRU HSE Moscow

1317 and 1318. sredin seminar

on Zoom, March 2 and 16, 2022

Outline

1. Signed two-mode networks
2. Salton's cosine
3. Jaccard's similarity
4. Remarks
5. References



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

Current version of slides (March 16, 2022 at 17 : 47): slides PDF

https://github.com/bavla/NormNet/tree/main/docs

Signed two-mode networks

A *signed two-mode network* is described by an affiliation matrix A with rows corresponding to *respondents*/actors U (primary mode) and columns corresponding to different *cases*/choices V (secondary mode).

The matrix entry $A[u, v]$ contains the decision/opinion of the respondent u about the case v and can take one of the values: p (Yes, Positive, Support), n (No, Negative, Oppose), a (Ambivalent, No opinion), r (Not valid), and z (NA, Not available, Absent) [3, 7].

For example, UN or parliament votings (U – countries / representatives; V – resolutions / bills); pre-election pools (U – respondents, V – selected politicians), etc.

The network matrix A depends on values p, n, a, r, z – we write $A(p, n, a, r, z)$. The notation $A(1, -1, 0, 0, 0)$ denotes a matrix obtained from A by substitutions $p \to 1$, $n \to -1$, $a \to 0$, $r \to 0, z \to 0$.

Supreme court

```
Supreme Court
Mrvar, A., Doreian, P. (2009) Partitioning Signed Two-Mode Networks.
Journal of Mathematical Sociology 33(3):196-221.
Case;  Alito; Breyer; Ginsburg; Kennedy; Roberts; Scalia; Souter; Stevens; Thomas
Abortion Rights             ; p; n; n; p; p; p; n; n; p
Anti Trust Law              ; p; p; p; z; p; p; p; p; n
Appelate Procedures         ; p; n; n; p; p; p; n; n; p
Bankruptcy Rules            ; n; p; p; p; n; n; p; p; n
Campaign Finance            ; p; n; n; p; p; p; n; n; p
Church and State            ; p; n; n; p; p; p; n; n; p
Court Sentencing            ; p; p; p; p; p; p; n; p; p
Death Penalty               ; p; n; n; p; p; p; n; n; p
Death Penalty 1             ; p; n; n; p; p; p; n; n; p
Death Penalty 2             ; p; n; n; p; p; p; n; n; p
Death Penalty Appeal        ; n; p; p; p; p; n; p; p; n
Disabled Students           ; p; p; p; p; p; n; p; p; n
EPA Pollution Control       ; n; p; p; p; n; n; p; p; n
Endengered Species          ; p; n; n; p; p; p; n; n; p
Energy Price Fixing         ; p; n; p; p; p; p; p; n; p
Equal Pay Rights            ; p; n; n; p; p; p; n; n; p
False Claims Act            ; p; z; n; p; p; p; p; n; p
Federal Employees           ; p; p; p; p; p; n; p; p; n
First Amendment             ; p; n; n; p; p; p; n; n; p
Grazing Rights              ; p; p; n; p; p; p; p; n; p
Illegal Searches            ; p; p; p; p; p; p; n; p; p
Immigration Drug Law        ; p; p; p; p; p; p; p; p; n
Interstate Telephones       ; p; p; p; p; p; n; p; p; n
Mental Health Death Penalty ; n; p; p; p; n; n; p; p; n
Patent Law                  ; p; p; p; p; p; p; p; p; n
Patents and Software        ; p; p; p; p; z; p; p; n; p
Police Discretion           ; p; p; p; p; p; p; p; n; p
Price Agreements            ; p; n; n; p; p; p; n; n; z
Prisoner Health             ; p; p; p; p; p; n; p; p; n
Punitive Damages            ; p; p; n; p; p; n; p; n; n
School Integration          ; p; n; n; p; p; p; n; n; p
Securities Fraud            ; p; p; p; p; p; p; p; n; p
Sentencing                  ; p; n; n; p; p; p; n; n; p
Sentencing 1                ; p; p; n; p; p; n; p; n; n
Texas Death Penalty 1       ; n; p; p; p; n; n; p; p; n
Texas Death Penalty 2       ; n; p; p; p; n; n; p; p; n
Texas Death Penalty 3       ; n; p; p; p; n; n; p; p; n
Trade Restraint             ; p; p; n; p; p; p; p; n; p
UN Property Taxes           ; p; n; p; p; p; p; p; n; z
Wrongful Arrest             ; p; n; n; p; p; p; p; p; p
```

[4]

Signed two-mode networks in R

```
> wdir <- ".../sreda/1317/SC"
> setwd(wdir)

> recode <- function(T,p,n,a,r,z){
+   A <- matrix(NA,nrow=nrow(T),ncol=ncol(T))
+   dimnames(A) <- list(rownames(T),colnames(T))
+   A[T=="p"] <- p; A[T=="n"] <- n; A[T=="a"] <- a;
+   A[T=="r"] <- r; A[T=="z"] <- z
+   return(A)
+ }

> T <- read.csv("SCa.txt",sep=";",head=TRUE,row.names=1,skip=3,
+   strip.white=TRUE,stringsAsFactors=FALSE)
> dim(T)
> L <- c("z","r","a","n","p")
> for(i in 1:9) T[,i] <- factor(T[,i],levels=L)
> head(T)

> A <- recode(T,1,2,3,4,5)
> head(A)
> A <- recode(T,1,-1,0,0,0)
> head(A)
```

Some datasets

1. **US senate and congress**
   Bradley Robinson US Senate Voting Records for 1990-2016 / Sessions 101-114.
   https://data.world/bradrobinson/us-senate-voting-records
   Bavla

2. **Bundestag**
   Data from https://github.com/jmbh/bundestag used by Jonas Haslbeck in Analyzing voting pattern of German parliament.
   Data for the time period 26.11.2014 - 14.04.2016 were obtained from: https://www.bundestag.de/bundestag/plenum/abstimmung
   Bavla

3. **Parlameter**
   Voting in Slovenian parliament https://parlameter.si/
   Bavla

Salton's cosine

Let us have data on the survey collected in the matrix (table) A of dimension $n \times m$
$a[k, i]$ = respondent k answer to the i-th case.

All the cases are measured on the same numerical scale. We also allow the value `NA` (not measured).

The usual *scalar product* of the vectors x and y is determined by the rule $x \bullet y = \sum_i x_i \cdot y_i$. The *length* of the vector x is $|x| = \sqrt{x \bullet x}$.

Weighted scalar product $x \odot y = \sum_i w_i \cdot x_i \cdot y_i$ can be introduced for a given vector of nonnegative weights w. The *weighted length* of the vector x is $\| x \| = \sqrt{x \odot x}$.

Let A be the matrix of a two-mode network and $W = A \bullet \mathrm{diag}(\sqrt{w}) = [\sqrt{w_i} \cdot a_{i,j}]$. Then $A^T \odot A = W^T \bullet W$ holds.

Salton's cosine

Salton's index $S(i,j)$ or cosine similarity of cases i and j is defined as the cosine between the vectors / variables $A[.,i]$ and $A[.,j]$

$$S(i,j) = \cos(A[.,i], A[.,j]) = \frac{\sum_k a[k,i] \cdot a[k,j]}{\sqrt{\sum_k a[k,i]^2 \cdot \sum_k a[k,j]^2}} = \frac{A[.,i] \bullet A[.,j]}{|A[.,i]| \cdot |A[.,j]|}$$

In general, $S(i,j) \in [-1,1]$ and $S(i,i) = 1$. $S(x, a \cdot x) = 1$ for $a > 0$.

The Salton's index can be transformed into dissimilarity (for clustering) in several ways:

$$d(i,j) = \frac{1 - S(i,j)}{2}$$

$$\delta(i,j) = \frac{\arccos(S(i,j))}{\pi}$$

Computing the Salton's similarity matrix

From matrix A we calculate (projection to cases) a square *matching matrix* of cases (more precisely, answers to questions)

$$C = A^T \bullet A$$

of dimensions $m \times m$. Its entry

$$c[i,j] = \sum_k a^T[i,k] \cdot a[k,j] = \sum_k a[k,i] \cdot a[k,j]$$

is equal to the scalar product in the numerator of Salton's similarity. More, $c[i,i] = \sum_k a[k,i] \cdot a[k,i] = \sum_k a[k,i]^2$ is the same as the term appearing in the denominator. Thus, Salton's similarity can be expressed in terms of the matching matrix

$$S(i,j) = \frac{c[i,j]}{\sqrt{c[i,i] \cdot c[j,j]}}$$

Salton's similarity for (p,n,a,r,z)-vectors

Let's look at a special case where only three values are possible: p, n, and $a = r = z$. The data is recoded as follows: $p \to 1$, $n \to -1$, $z \to 0$. Let $A_s = A(1, -1, 0, 0, 0)$ contains the recoded data. Then from the form for $c[i,j]$ it follows

$$c[i,j] = \#(1,1|-1,-1) - \#(1,-1|-1,1) =$$

$$= \#\text{consistent answers} - \#\text{opposite answers}$$

and

$$c[i,i] = \#\text{answers to question } i$$

Thus, $S(i,j)$ is equal to the consistency of the answers divided (normalized) by the geometric mean of the numbers of answers to both questions.

We can relatively easily show that $S(i,j) = 1 \Leftrightarrow a[.,i] = a[.,j]$ and $S(i,j) = -1 \Leftrightarrow a[.,i] = -a[.,j]$.

US Senate 106/1 dendrogram

Signed projections

V. Batagelj

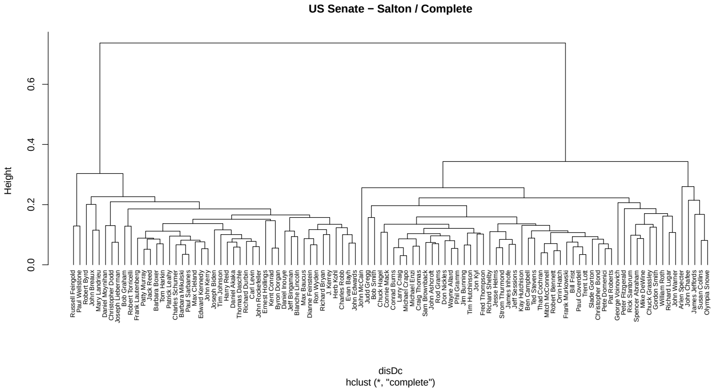
Signed two-mode networks

Salton's cosine

Jaccard's similarity

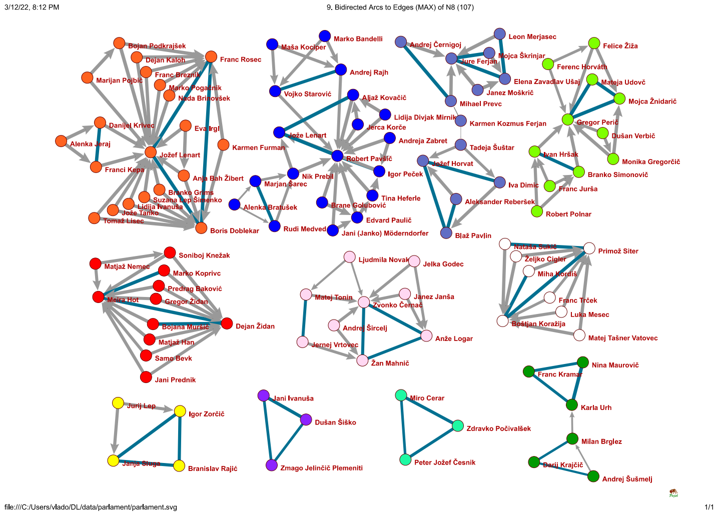
Remarks

References



Slovenian parliament dendrogram

Slovenian parliament 2 most similar neighbors network



Jaccard's similarity

Jaccard's similarity can be generalized to the scale (p,n,z) as follows: By recoding, we create the matrices $A_p = A(1,0,0,0,0)$ and $A_n = A(0,1,0,0,0)$ and calculate the matching matrix

$$C = A_p^T \bullet A_p + A_n^T \bullet A_n$$

with entries

$$c[i,j] = \#(1,1|-1,-1) = \#\text{consistent answers}$$

and $c[i,i] = \#$answers to question i.
Therefore Jaccard's similarity ($J = \frac{|X \cap Y|}{|X \cup Y|}$)

$$J(i,j) = \frac{c[i,j]}{c[i,i] + c[j,j] - c[i,j]}$$

and the Jaccard dissimilarity $d_J(i,j) = 1 - J(i,j)$

Jaccard's similarity

There are some generalizations of the Jaccard similarity $J(x, y)$ to (nonnegative) real vectors. Perhaps the most famous is (Ružička)

$$J_1(x, y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

We can easily check that for binary vectors x and y hold $\sum_i \min(x_i, y_i) = |X \cap Y|$ and $\sum_i \max(x_i, y_i) = |X \cup Y|$. So, for binary vectors we $J_1(x, y) = J(x, y)$.

Unfortunately, this measure does not work well for vectors containing negative values.
Another generalization of Jaccard's similarity is

$$J_2(x, y) = \frac{x \bullet y}{(x - y)^2 + x \bullet y}, \quad \text{for } x \neq 0 \text{ and } y \neq 0$$

and $J_2(0, 0) = 1$.

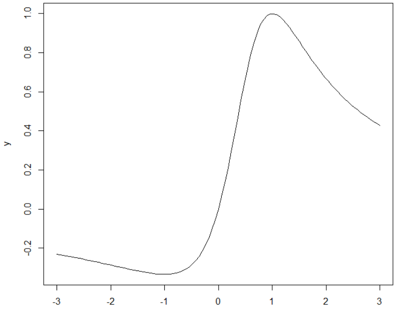
Again, for binary vectors x and y we have $x \bullet y = |X \cap Y|$ and $(x - y)^2 + x \bullet y = |X \cup Y|$.

Jaccard's similarity

The following properties hold for general real vectors:

- $J_2(x, y) = J_2(y, x)$
- $J_2(x, y) \leq 1$
- $J_2(x, y) = 1 \Leftrightarrow x = y$
- $J_2(a \cdot x, a \cdot y) = J_2(x, y)$ for $a \neq 0$.

For nonnegative vectors x and y, $J_2(x, y) \geq 0$. But in the case of $y = -x$ we get $J_2(x, -x) = -\frac{1}{3}$. If we look a little more generally, we get $J_2(x, a.x) = \frac{a}{(1-a)^2+a}$.



Jaccard's similarity

$-\frac{1}{3}$ seems to be the smallest possible value of measure J_2.
That turns out to be true. The proof is short: the inequality $0 \leq (x+y)^2$ is transformed into $-3 \cdot x \bullet y \leq (x-y)^2 + x \bullet y$ which means $J_2(x,y) \geq -\frac{1}{3}$; the equality holds iff $x = -y$.
Measure J_2 is maped to the interval [1,0] by transformation

$$J_3(x,y) = \frac{3 \cdot J_2(x,y) + 1}{4}.$$

If we express (simplify) J_3 with scalar products, we get

$$J_3(x,y) = \frac{(\frac{x+y}{2})^2}{(x-y)^2 + x \bullet y}$$

Generalized binary similarities

Signed projections

V. Batagelj

Signed two-mode networks

Salton's cosine

Jaccard's similarity

Remarks

References

Most of similarities between binary vectors are defined/expressed using matching counters $a = \#(1,1)$, $b = \#(1,0)$, $c = \#(0,1)$ and $d = \#(0,0)$. This approach can be extended also to signed data. To obtain the counters used in a selected similarity we can use the approach we used for extending the Jaccard's similarity.
For example, to get counters $pn = \#(p,n)$ we create matrices $A_p = A(1,0,0,0,0)$ and $A_n = A(0,1,0,0,0)$ and compute their product

$$C_{pn} = A_p^T \bullet A_n$$

with entries

$$c_{pn}[i,j] = \#(p,n) = \#\text{consistent answers}$$

and $c_{pn}[i,i] = 0$. Note that, since $C_{np} = C_{pn}^T$

$$c_{pn}[i,j] + c_{pn}[j,i] = \#(p,n|n,p) = \#\text{opposite answers}$$

Manhattan-Jaccard's dissimilarity

I also tried the dissimilarity that uses the Jaccard's normalization on the Manhattan distance

$$MJ(x, y) = \frac{\sum_i w_i \cdot |x_i - y_i|}{|X \cup Y|}$$

where $|X \cup Y| = \sum_i \max(|x_i|, |y_i|)$.

Remarks

1. Generalization of some other (dis)similarities for binary vectors to (p, n, a, r, z)-vectors [5].
2. US senate, Bundestag, Parlameter. Additional data: parties, short names of cases, keywords/tags, types of cases: poslanci, besede, ki so zaznamovale sejo
3. Schoch [7], complex numbers.
4. semirings [3].

References

- Batagelj, V.: On fractional approach to analysis of linked networks. Scientometrics 123 (2020) 2: 621-633
- Batagelj, V.: Analysis of the Southern women network using fractional approach. Social Networks 68(2022), 229-236
- Batagelj, V. (1994). Semirings for social networks analysis. The Journal of Mathematical Sociology, 19(1), 53–68.
- Mrvar, A., Doreian, P. (2009) Partitioning Signed Two-Mode Networks. Journal of Mathematical Sociology 33(3):196-221.
- Deza, M.M., Deza, E. (2013). Encyclopedia of Distances. Springer-Verlag, Berlin Heidelberg.
- Robinson, B. (2022) US Senate Voting Records https://data.world/bradrobinson/us-senate-voting-records/activity
- Schoch, D. (2021). Projecting signed two-mode networks, The Journal of Mathematical Sociology, 45:1, 37-50.