OpenAlex2Pajek

Institutions, the saturation approach, and co-authorship between countries

Vladimir Batagelj
IMFM, UP IAM

1351. + 1352. sredin seminar
Ljubljana, May 22 and 29, 2024

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (May 29, 2024 at 05 : 36): slides PDF
https://github.com/bavla/OpenAlex

OpenAlex2Pajek

We continue the development of support for the conversion of OpenAlex data into Pajek's networks.

The saturation approach was split into two phases:

- the saturation phase dealing only with the citation network for the selection of the set of relevant works W
- creation of bibliographic networks for the selected set of relevant works W

OpenAlex2Pajek

saturation approach

The set W is determined iteratively using the function `OpenAlex2PajekCite`.

1. Create the basic query Q and determine using `OpenAlex2PajekCite` the initial version of W; list of old candidates C is empty
2. Analyze using Pajek macro `expNodes` the obtained citation network and identify new candidates N for relevant works. If N is empty **STOP**.
3. Save the list N in a CSV file. Using in R the command `joinLists("Cold.csv","N.csv","Cnew.csv")` join the old candidates and new candidates into the current list of candidates (removing duplicates).
4. Using `OpenAlex2PajekCite` determine the new version of W; go to **2**.

Creating a collection

OpenAlex2Pajek

saturation approach

To create the collection we first change the parameter `select` in the query Q to `selAll`. Afterward, we run the function `OpenAlex2PajekAll`.

Currently, we get a collection of bibliometric networks:

```
>>> n Citation Cite
>>> c publication year
>>> c type of publication
>>> c language of publication
>>> c cited by count
>>> c countries distinct count
>>> c referenced works
>>> n Authorship WA
>>> n Sources WJ
>>> n Keywords WK
>>> n Countries WC
```

and additionally names of works `xyzW.nam` and names of authors `xyzA.nam`.

OpenAlex2Pajek

saturation approach

I don't like the keywords provided by OpenAlex. In a future version of OpenAlex2Pajek I will provide an alternative based on words from the work's title (and abstract).

In phase one we could consider also other available properties of nodes (works).

On the **to do** list is to remove the use of Pajek from phase one and program the iterations in R.

OpenAlex2Pajek

converting dictionary into data frame

Internal: dict2DF

```
dict2DF <- function(dict,ind) {
  DF <- as.data.frame(do.call(rbind, as.list(dict)))
  return(DF[order(unlist(unname(DF[[ind]]))),])
}
```

OpenAlex2Pajek

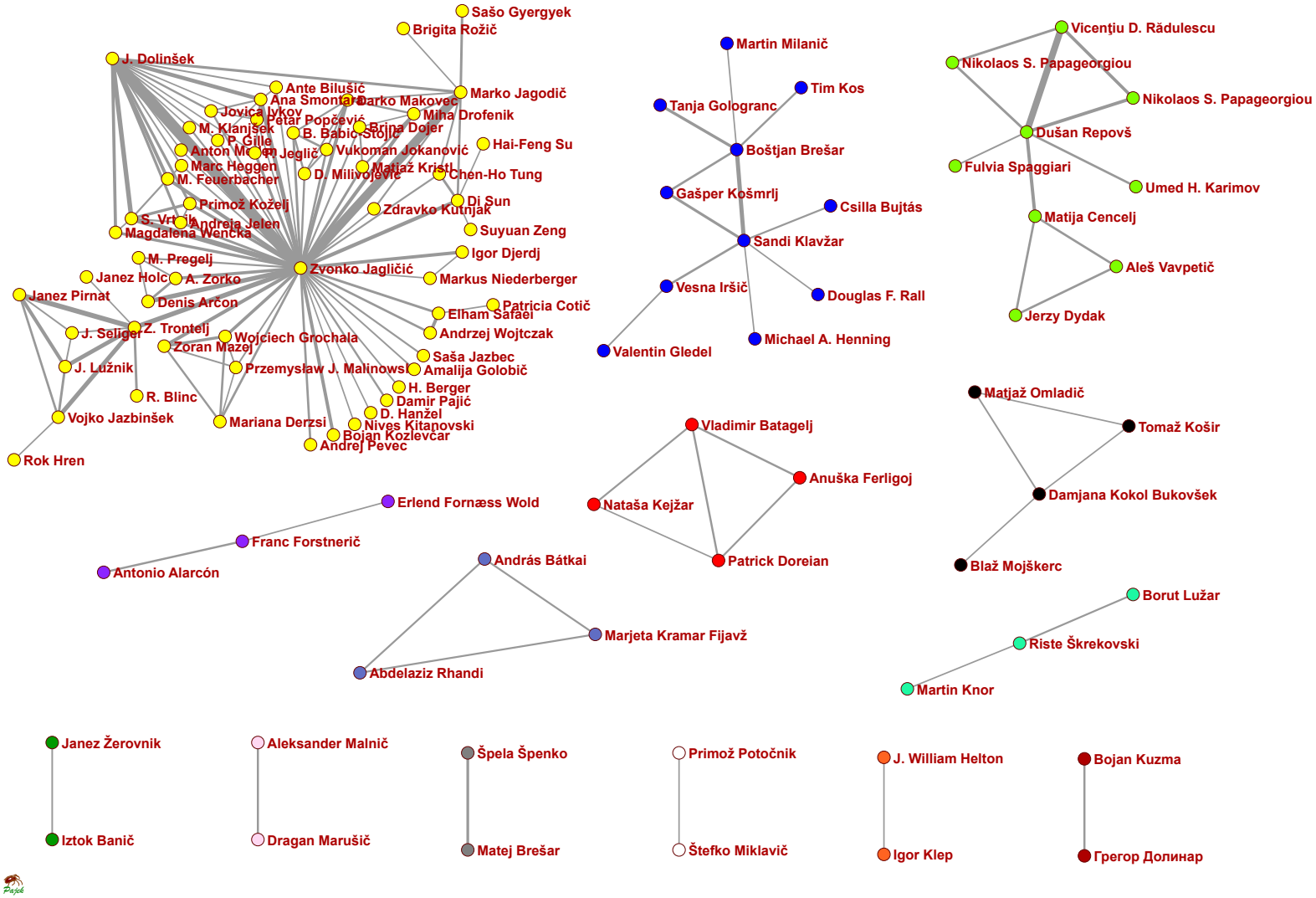
Institutions

In some cases, such as all works of researchers from a selected institution, the saturation phase is not needed.

Internal: Young universities
GitHub: HKUST, IMFM

OpenAlex2Pajek

IMFM co-authorship link cut at level 7



OpenAlex2Pajek

Co-authorship between countries

We developed a function `coAuthorship` that creates a sequence of temporal networks describing the co-authorship between world countries.

Internal: meaning; total; years

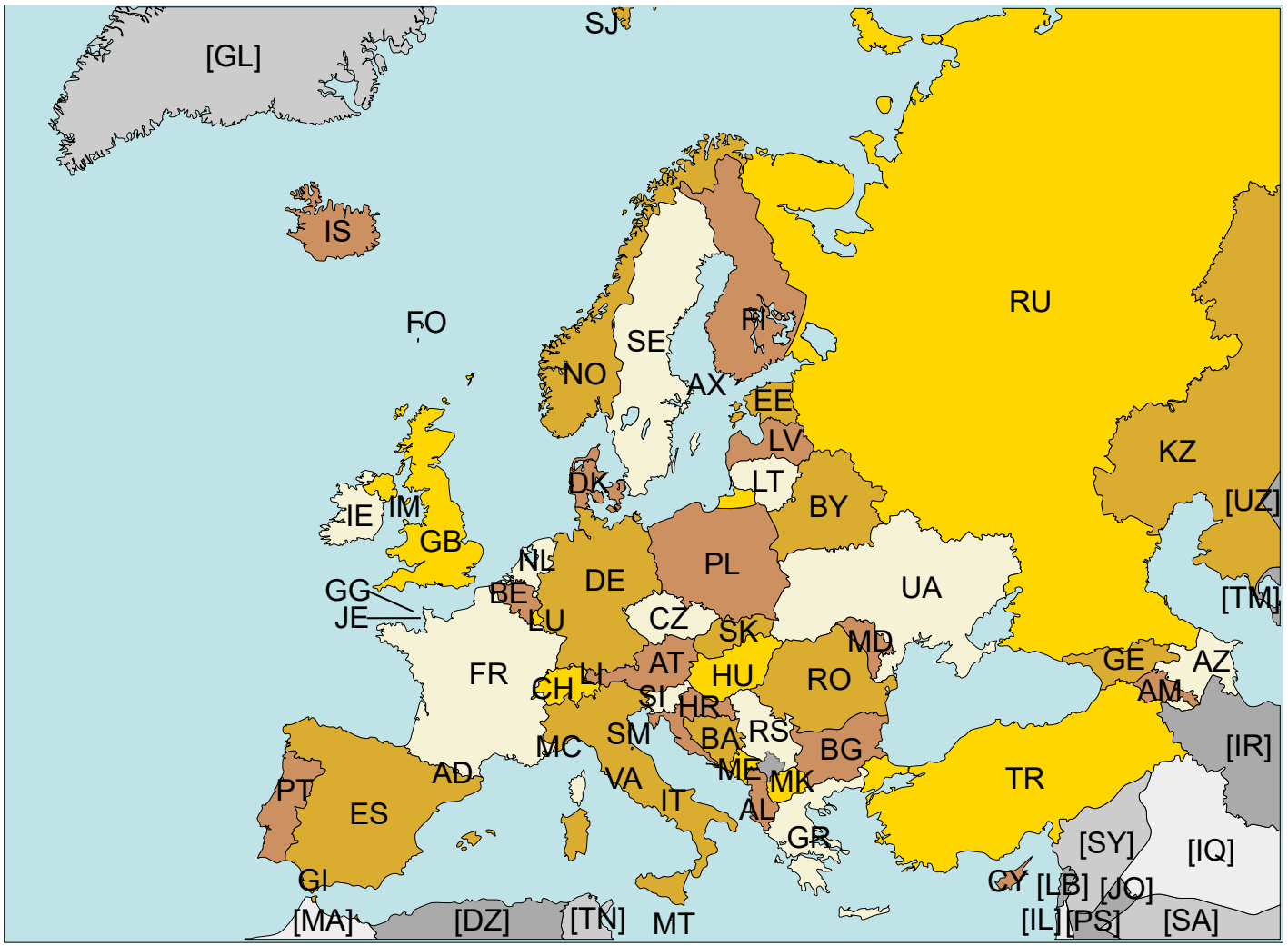
GitHub/Bavla/OpenAlex: pics; world, 1-neighbors PDF, Europe, 1-neighbors

Problem: OpenAlex is using ISO 2-character country codes. Only currently existing countries are considered.

It seems that OpenAlex exports data for only up to 200 most active countries.
Assuming the symmetry of the countries' co-authorship matrix we can get a complete matrix.

OpenAlex2Pajek

European countries



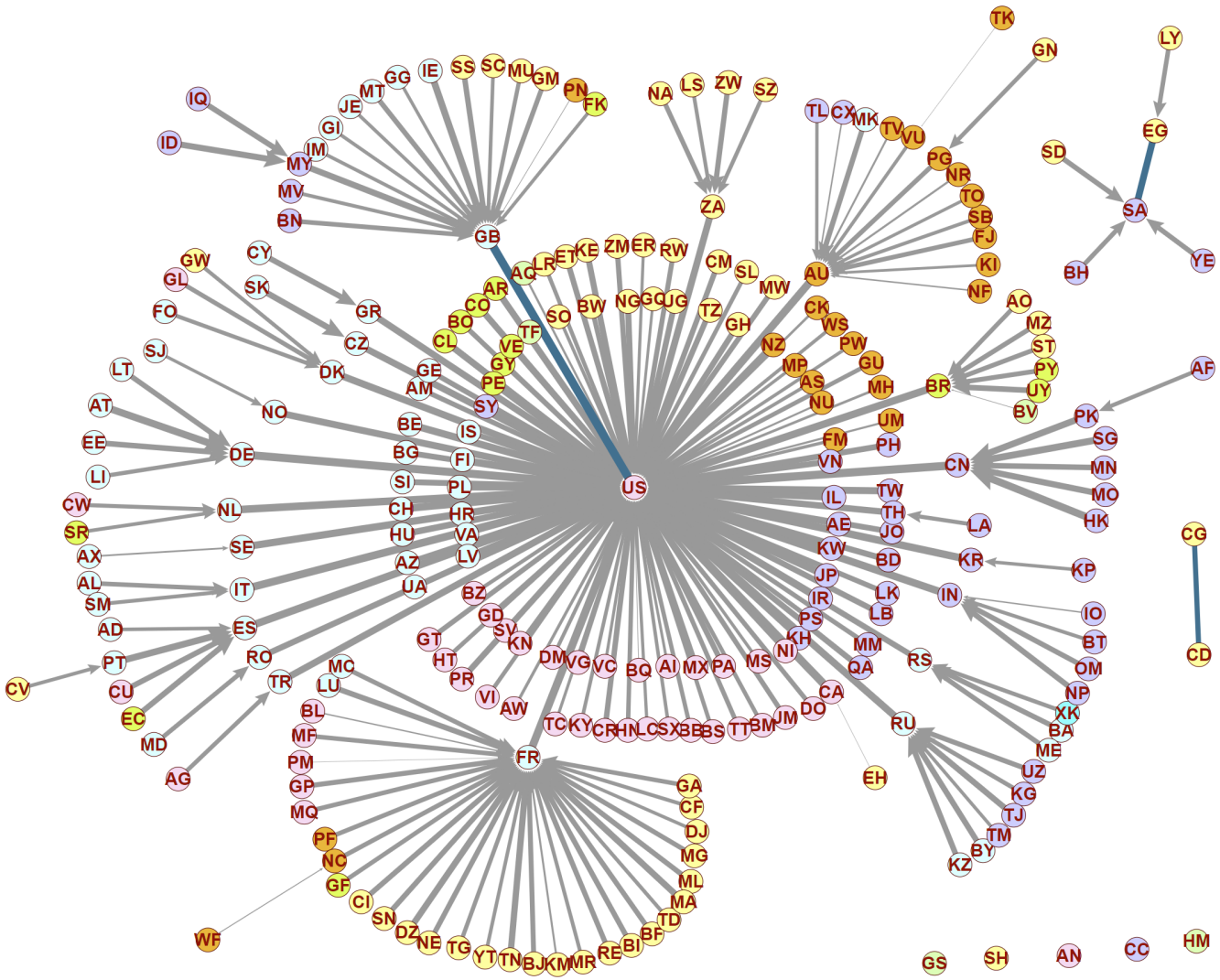
OpenAlex2Pajek

European countries

ISO	country	ISO	country	ISO	country
AD	Andorra	GB	G Britain	MK	N Macedonia
AL	Albania	GE	Georgia	MT	Malta
AM	Armenia	GG	Guernsey/GB	NL	Netherlands
AT	Austria	GI	Gibraltar/GB	NO	Norway
AX	Åland/FI	GR	Greece	PL	Poland
AZ	Azerbaijan	HR	Croatia	PT	Portugal
BA	Bosnia+Herz	HU	Hungary	RO	Romania
BE	Belgium	IE	Ireland	RS	Serbia
BG	Bulgaria	IM	i of Man/GB	RU	Russia
BY	Belarus	IS	Iceland	SE	Sweden
CH	Switzerland	IT	Italy	SI	Slovenia
CY	Cyprus	JE	Jersey/GB	SJ	Svalbard+JM
CZ	Czech rep	KZ	Kazakhstan	SK	Slovakia
DE	Germany	LI	Liechtenstein	SM	San Marino
DK	Denmark	LT	Lithuania	TR	Turkey
EE	Estonia	LU	Luxembourg	UA	Ukraine
ES	Spain	LV	Latvia	VA	Vatican
FI	Finland	MC	Monaco	XK	Kosovo
FO	Faroe i/DK	MD	Moldova		
FR	France	ME	Montenegro		

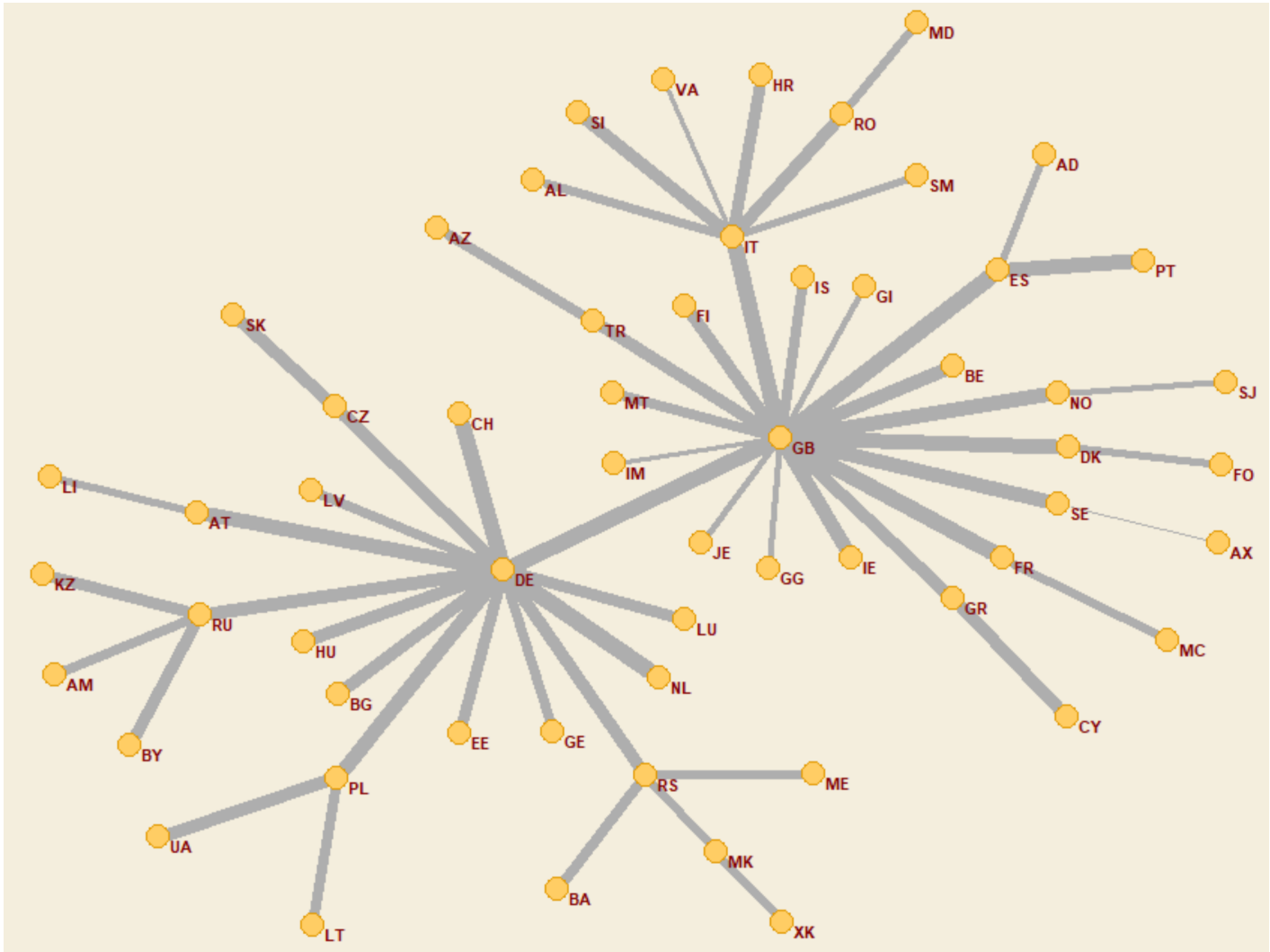
OpenAlex2Pajek

Total co-authorship between world countries/1-neighbors



OpenAlex2Pajek

Co-authorship between European countries 2020 /1-neighbors



Internal: Europe; GitHub: Europe

OpenAlex2Pajek
Corrected Euclidean distance

When computing dissimilarity $D[a, b]$ between nodes a and b in a co-authorship network it is important to use the corrected dissimilarities. We selected the corrected Euclidean distance [1]

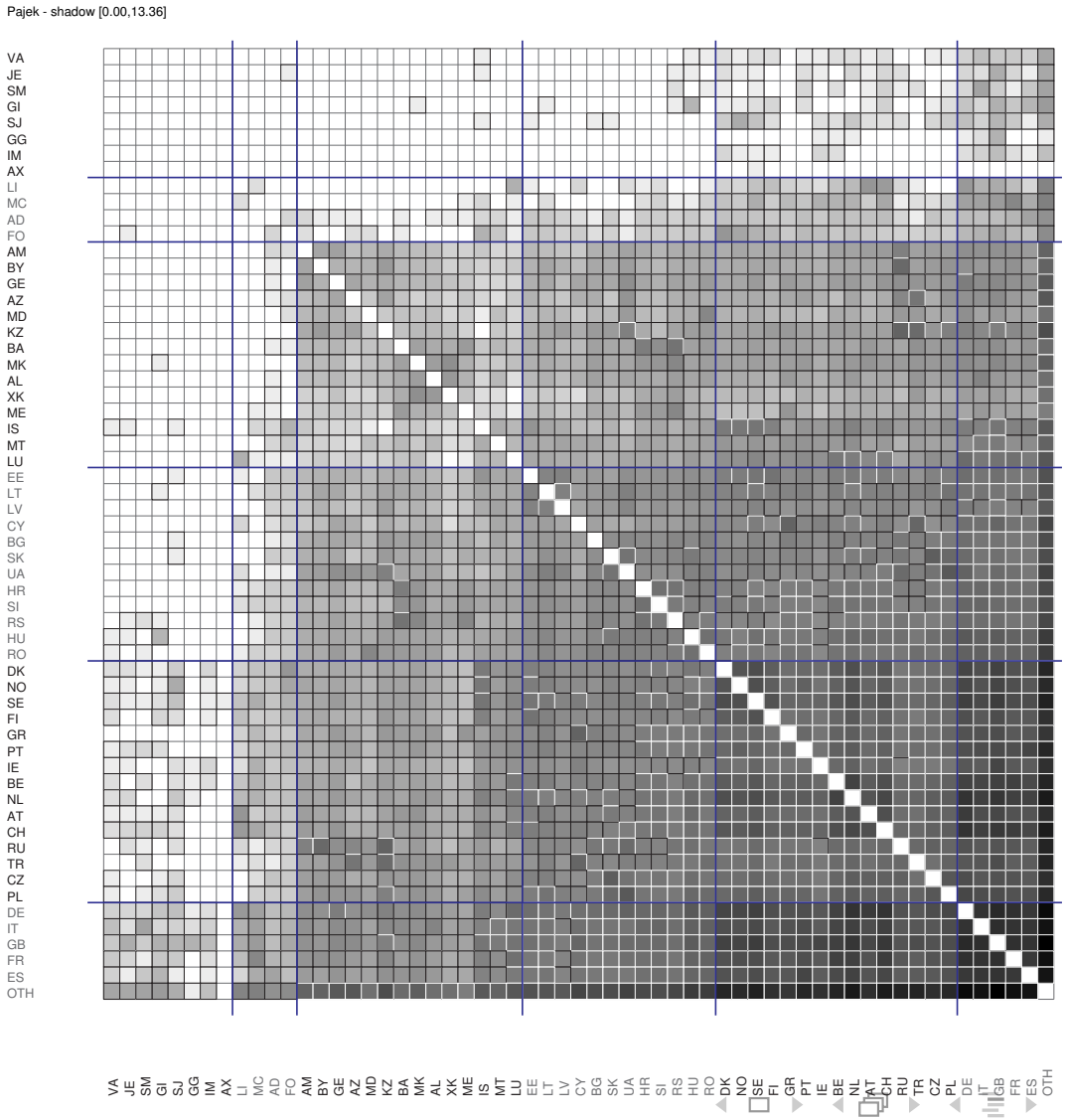
$$D[a,b] = \sqrt{(C[a,b]-C[b,a])^2 + (C[a,a]-C[b,b])^2 + \sum_{c:c\neq a, c\neq b} (C[a,c]-C[b,c])^2}$$

For clustering co-authorship networks we transformed the weights using $w' = 1 + \ln(w)$ – a balance between the structure (links) and large range of weights. Also convenient for visualization.

GitHub/Bavla/OpenAlex: Clustering

OpenAlex2Pajek

European countries 2023 / Clustering



OpenAlex2Pajek

Normalizations: Affinity

The co-authorship network is represented by a square matrix C on the set of countries. Its entry $C[a, b]$ is equal to the number of co-authorships between countries a and b in a selected time period. Non-existing links are represented with the value 0. The matrix C is symmetric. Let us denote the row sum $R(a) = \sum_b C[a, b]$ [2].

Affinity (Stochastic, Markov, Output, row) normalization
For $R(a) > 0$ [3, p. 631]

$$M[a, b] = \frac{C[a, b]}{R(a)}$$

If $R(a) = 0$ hen also $M[a, b] = 0$.
For $R(a) > 0$, we have $\sum_b M[a, b] = 1$. The matrix entries $M[a, b]$ can be interpreted as probabilities that an author from the country a collaborates with an author from country b.

OpenAlex2Pajek

Normalizations: Jaccard and Salton

Jaccard

$$J[a,b] = \frac{C[a,b]}{R(a) + R(b) - C[a,b]}$$

Salton (cosine)

$$S[a,b] = \frac{C[a,b]}{\sqrt{R(a) \cdot R(b)}}$$

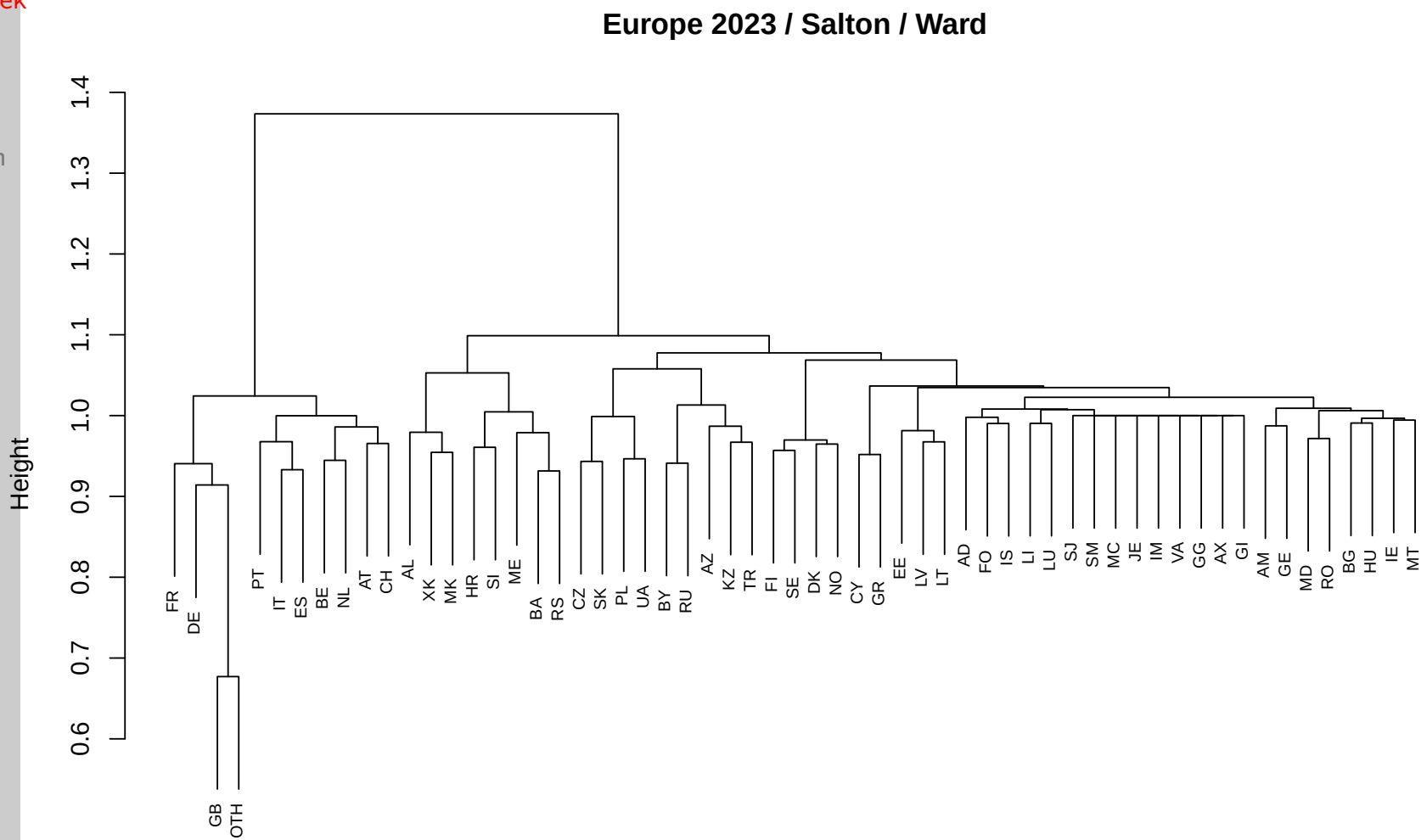
OpenAlex2Pajek

Normalizations / Salton and Jaccard

```
> i <- 34
> P <- E[[i]]$M; diag(P) <- 0; diag(P) <- rowSums(P)
> S <- P; diag(S) <- 1; J <- P; diag(J) <- 1
> n = nrow(S)
> for(u in 1:(n-1)) for(v in (u+1):n) {
+    S[v,u] <- S[u,v] <- P[u,v]/sqrt(P[u,u]*P[v,v])
+    J[v,u] <- J[u,v] <- P[u,v]/(P[u,u]+P[v,v]-P[u,v]) }
> matrix2net(S,Net="EuSalton2023.net")
> matrix2net(J,Net="EuJaccard2023.net")
> DS <- as.dist(1-S)
> t <- hclust(DS,method="ward.D")
> plot(t,hang=0.2,main="Europe 2023 / Salton / Ward",cex=0.7)
> DJ <- as.dist(1-J)
> h <- hclust(DJ,method="ward.D")
> plot(h,hang=0.2,main="Europe 2023 / Jaccard / Ward",cex=0.7)
```

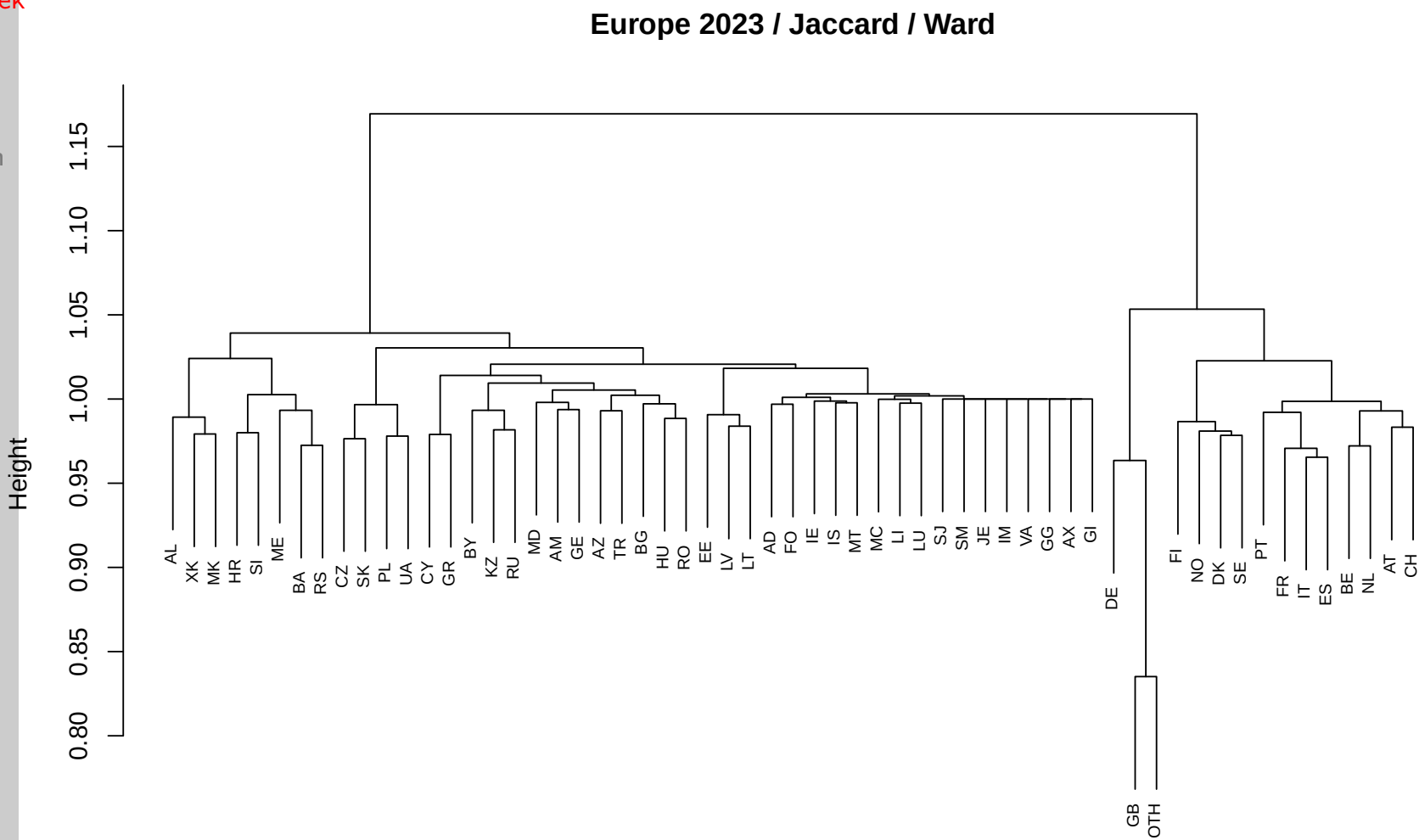
OpenAlex2Pajek

European countries 2023 / Salton



OpenAlex2Pajek

European countries 2023 / Jaccard



OpenAlex2Pajek

Normalizations: Activity (Balassa)

Let $Q(a) = \sum_b C[b, a]$ denote the column sum for the country a, and $T = \sum_{a,b} C[a, b]$ the total sum of weights in the network. If our network $R(a) = Q(a)$. Then $R(a)/T$ is the probability of activity of country a. The expected weight $E[a, b]$ from a to b is equal to:

$$E[a, b] = \frac{R(a)}{T} \cdot Q(b)$$

The measured weight $C[a, b]$ may deviate by a factor $A(a, b)$ from the expected value, $C[a, b] = A[a, b] \cdot E[a, b]$, or [3, p. 633]

$$A(a, b) = \frac{C[a, b] \cdot T}{R(a) \cdot Q(b)}$$

If $A[a, b] > 1$ the measured weight is larger than expected.

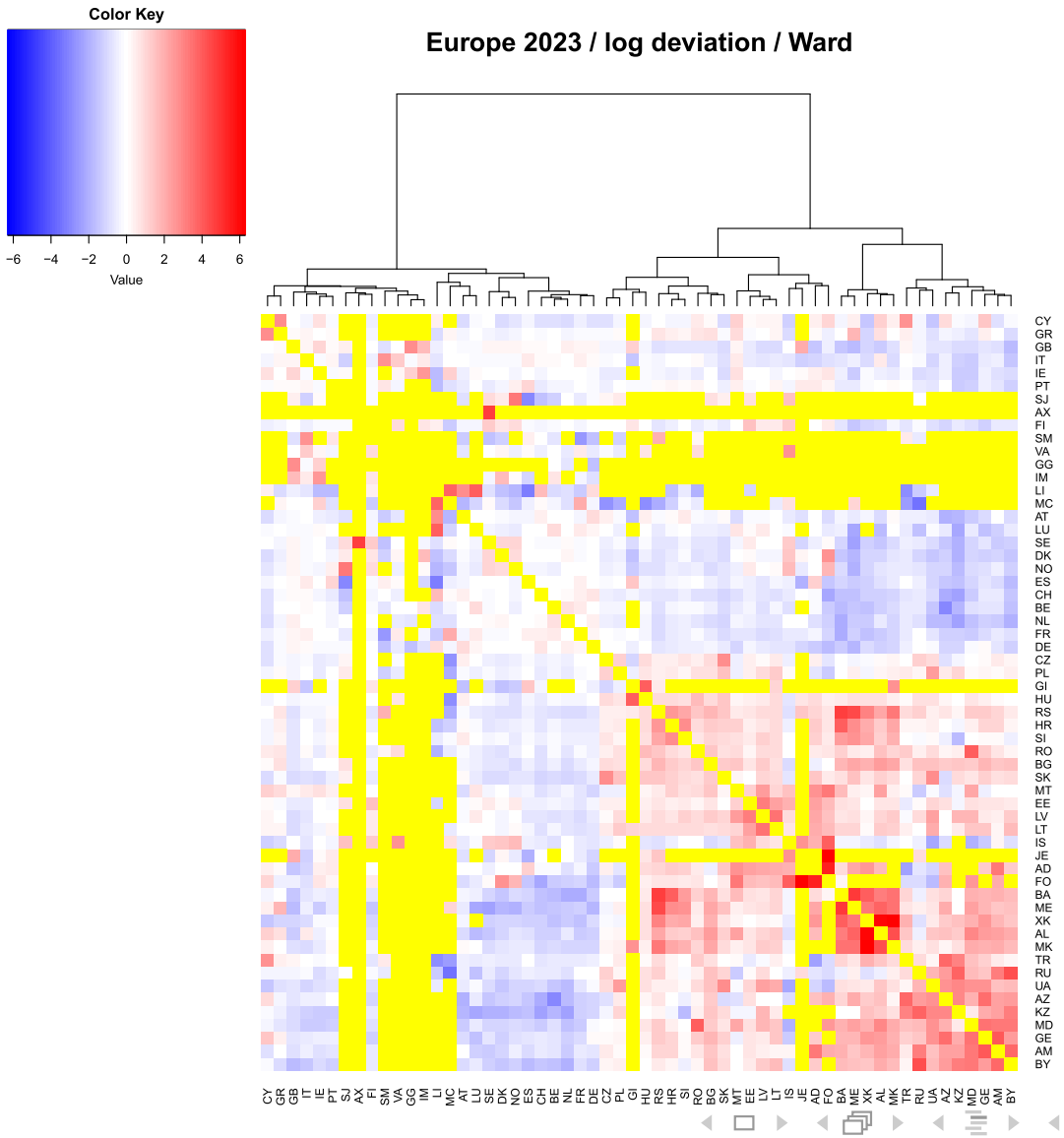
OpenAlex2Pajek

Normalizations: Activity (Balassa)

The deviation measure A is called the activity index (also Balassa index or the "revealed comparative advantage" [4]). The range of A is not 'symmetric'. We apply a logarithmic function to it [5]. For easier interpretation, we selected base 2 logarithms:

$$B[a, b] = \log_2 A[a, b] \quad \text{for} \quad A[a, b] > 0$$

If B[a, b] = 0, the collaboration is equal to the expected value. In our analysis we used the index B. We have A[a,b] = 0 for non-linked countries. We set B[a,b] = 0 in such cases.



OpenAlex2Pajek
Conclusions

1. OpenAlex is a rich source of bibliometric data relatively easy to use.
2. Local copy of OpenAlex !?
3. Developement of higher order bibliographic services.
4. Analyses of interesting bibliographies.

Acknowledgments

The computational work reported in this paper was performed using a collection of R functions `OpenAlex`, R program `OpenAlex2Pajek`, and the program Pajek for analysis of large networks. The code and data are available at Github/Bavla/OpenAlex.

This work is supported in part by the Slovenian Research Agency (research program P1-0294, research program CogniCom (0013103) at the University of Primorska, and research projects J5-2557, J1-2481, and J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc).

References I

1. Doreian, P., Batagelj, V. & Ferligoj, A. *Generalized blockmodeling*. **25** (Cambridge university press, 2005).
2. Matveeva, N., Batagelj, V. & Ferligoj, A. Scientific collaboration of post-Soviet countries: the effects of different network normalizations. *Scientometrics* **128,** 4219–4242 (2023).
3. Zitt, M., Bassecoulard, E. & Okubo, Y. Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics* **47,** 627–657 (2000).
4. Balassa, B. Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School* **33,** 99–123 (1965).
5. Vollrath, T. L. A theoretical evaluation of alternative trade intensity measures of revealed comparative advantage. *Weltwirtschaftliches Archiv* **127,** 265–280 (1991).