Truncated two-mode network projections

Vladimir Batagelj
IMFM, UP IAM

INSNA Sunbelt 2024
Edinburgh, UK, June 24-30, 2024

Outline

redbubble

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (June 28, 2024 at 00 : 40): slides PDF
https://github.com/bavla/wNets

Collections of networks

From a selected bibliographic data we can construct a collection of corresponding bibliographic networks such as the authorship network $\mathcal{N}_{WA}$, the keywords network $\mathcal{N}_{WK}$, the citation network $\mathcal{N}_{Ci}$, etc. [1]. For example, in the two-mode authorship network $\mathcal{N}_{WA} = ((W, A), L_{WA})$ the set of nodes is split to the set of works W and the set of authors A. The set of links L_{WA} consists of arcs (directed links) $(w, a) \in L_{WA}$ stating that the work $w \in W$ was co-authored by the author $a \in A$. The citation network $\mathcal{N}_{Ci} = (W, L_{Ci})$ is a directed one-mode network on works with arcs $(w, z) \in L_{Ci}$ stating that the work $w \in W$ is citing the work $z \in W$.

Networks from a collection share some set of nodes. For example the authorship network $\mathcal{N}_{WA}$ and the citation network $\mathcal{N}_{Ci}$ share the set of works W. This allows us to compute, using network matrix multiplication, derived networks such as the authors$\times$ keywords network $\mathbf{AK} = \mathbf{WA}^T \cdot \mathbf{WK}$ or the network of citations between authors $\mathbf{ACiA} = \mathbf{WA}^T \cdot \mathbf{Ci} \cdot \mathbf{WA}$ ($\mathbf{WA}^T$ is the transpose of matrix $\mathbf{WA}$) .

Projections

A special type of derived networks are projections:

Co-appearance network $\mathbf{Co} = \mathbf{WA}^T \cdot \mathbf{WA}$
Row (out-)contribution network $\mathbf{Cr} = \mathbf{WA}^T \cdot \mathrm{bin}(\mathbf{WA})$
Normalized (fractional) co-appearance network
$\mathbf{Cn} = n(\mathbf{WA})^T \cdot n(\mathbf{WA})$
Strict (Newman's) co-appearance network
$\mathbf{Cn}' = D_0(n(\mathbf{WA})^T \cdot N(\mathbf{WA}))$, where $D_0(\mathbf{M})$ sets the diagonal of matrix $\mathbf{M}$ to 0.

The *standard normalization* matrix $n(\mathbf{WA}) = [wan[w,a]]$, where $wan[w,a] = \frac{wa[w,a]}{\mathrm{wod}_{WA}(w)}$ for $\mathrm{wod}_{WA}(w) > 0$.
For binary matrix, its *strict (Newman's) normalization* matrix $N(\mathbf{WA}) = [wan'[w,a]]$, where $wan'[w,a] = \frac{wa[w,a]}{\mathrm{od}_{WA}(w)-1}$ for $\mathrm{od}_{WA}(w) > 1$.

Outer product decomposition

The network product is determined by the product of corresponding matrices. Large networks are usually sparse – a graph representation is used to avoid computations with zeros.

For vectors $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and $\mathbf{y} = [y_1, y_2, \ldots, y_m]$ their *outer product* $\mathbf{x} \circ \mathbf{y}$ is defined as a matrix $\mathbf{x} \circ \mathbf{y} = [x_i \cdot y_j]_{n \times m}$. Then we can express the product $\mathbf{C}_{I \times J}$ of two compatible matrices $\mathbf{A}_{I \times K}$ and $\mathbf{B}_{K \times J}$ as the *outer product decomposition* [2]

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \sum_{k \in K} \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k = \mathbf{A}[\cdot, k] \circ \mathbf{B}[k, \cdot],$$

where $\mathbf{A}[\cdot, k]$ is the k-th column of matrix $\mathbf{A}$, and $\mathbf{B}[k, \cdot]$ is the k-th row of matrix $\mathbf{B}$.

The computational complexity of the network product is larger than $O(\Delta^2)$ and not larger than $O(|K| \cdot \Delta^2)$, Δ is the maximum degree in K.

The problem

Large bibliographic networks are sparse – the number of links is of the same order as the number of nodes (the average node degree is small). This is not necessarily true for their product – in some cases, it can "explode" (it is not sparse, increases in time and space complexity) [1].

The problem with space (computer memory) can be dealt with using a sparse matrix representation and partitioning matrices into blocks. An approach to deal with the time complexity is to reduce the size of the problem by limiting our attention to a selected subset of important nodes and computing with corresponding truncated networks.

The nodes can be selected by different criteria. An option is to consider the most important nodes in the derived network – nodes with the largest weighted degree. It turns out that the weighted degrees can be computed efficiently without computing the derived network itself.

iMetrics network and HKUST1 network

The *iMetrics* network was analyzed in the paper [3] and is an example of a typical authorship network.

The network *HKUST1* was created from Scopus by Nataliya Matveeva for her study of young universities [4]. It contains information about papers published in the years 2017-2019 by members of HKUST (The Hong Kong University of Science and Technology).

Its number of authors per paper distribution is a combination of a "regular" part (typical for authorship networks) and a large number of papers with more than 1000 co-authors – an "irregular" part. The paper on the ATLAS and CMS experiments [5] has 5215 co-authors. The second "largest" paper has 5096 co-authors, and there are 295 papers with the number of co-authors in the interval 2824–2953.

Truncated networks

V. Batagelj

Introduction

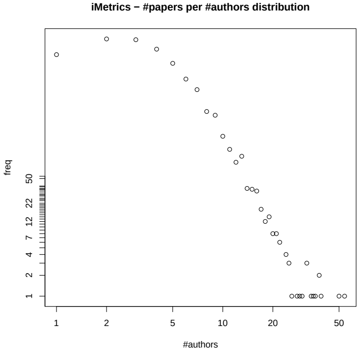
Truncated derived networks

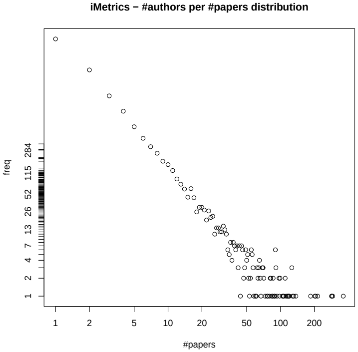
Example

Conclusions

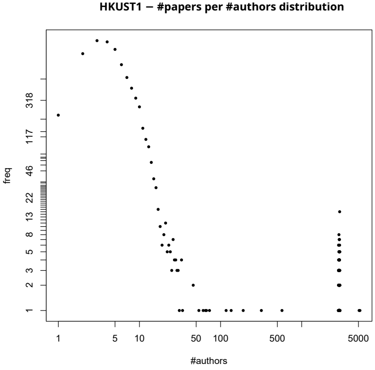
References

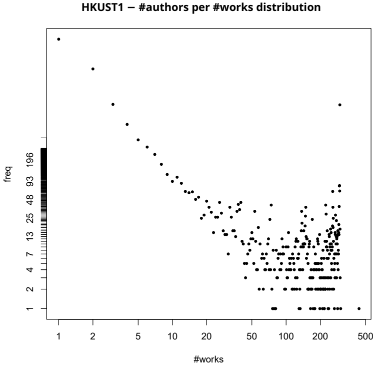
iMetrics distributions





HKUST1 distributions





iMetrics network and HKUST1 network

This is nothing special these days [6] – the Guinness World Record 653537 states: The most authors in a single peer-reviewed academic paper are 15025 and were achieved by the COVIDSurg and GlobalSurg Collaboratives at the University of Birmingham and the University of Edinburgh in the UK, as verified on 24 March 2021 [7].

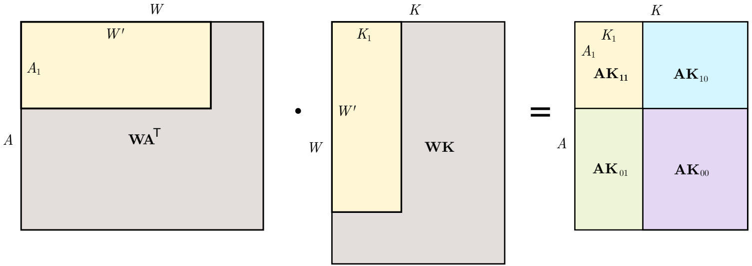
The problem is that such a paper contributes $15025^2 = 225\,750\,625$ links with a weight $4.429667\,10^{-9}$ to the network **Cn**.

Some months later they published two papers with 16162 co-authors.

For a network $\mathcal{N}$ with matrix $\mathbf{M}$, we define its *total* $T(\mathcal{N}) = T(\mathbf{M})$ as the sum of all its entries,

$$T(\mathcal{N}) = \sum_{e \in L} w(e) = \sum_{u \in U} \sum_{v \in V} m[u, v] = T(\mathbf{M})$$

Truncated derived networks



Let's split the set of authors A into two sets A_1 (selected authors) and A_0 (remaining authors), $A_1 \cup A_0 = A$ and $A_1 \cap A_0 = \emptyset$. We do the same with the set of keywords K – K_1 (selected keywords) and K_0 (remaining keywords), $K_1 \cup K_0 = K$ and $K_1 \cap K_0 = \emptyset$.

. . . Truncated derived networks

We call a *truncated derived network* the network

$$\mathbf{AK}_{11} = \mathbf{AK}[A_1, K_1] = \mathbf{WA}^T_{W' \times A_1} \cdot \mathbf{WK}_{W' \times K_1}$$

where $W' = \{w \in W : (\mathrm{od}_{WA_{W \times A_1}}(w) > 0) \wedge (\mathrm{od}_{WK_{W \times K_1}}(w) > 0)\}$.

For a selected author $a \in A_1$, we denote with $\mathrm{oInt}(a) = \mathrm{wod}_{AK_{11}}(a)$ her/his *internal* out-contribution, and with $\mathrm{oExt}(a) = \mathrm{wod}_{AK}(a) - \mathrm{oInt}(a)$ her/his *external* out-contribution.

And similarly, for a selected keyword $k \in K_1$,we denote with $\mathrm{iInt}(k) = \mathrm{wid}_{AK_{11}}(k)$ its *internal* in-contribution, and with $\mathrm{iExt}(k) = \mathrm{wid}_{AK}(k) - \mathrm{iInt}(k)$ its *external* in-contribution.

In the following, we shall show that weighted degrees of **AK** can be determined without computing the network itself.

. . . Truncated derived networks

We reorder the nodes of the network **AK** according to the A_1, A_0 and K_1, K_0 splits (see figure). The derived network matrix **AK** is split into four submatrices $\mathbf{AK}_{ij}, i,j \in \{0,1\}$. We denote their totals $T_{ij} = T(\mathbf{AK}_{ij})$. T_{11} is the contribution of cooperation among selected nodes, $T_{10} + T_{01}$ is the contribution of cooperation of selected nodes with remaining nodes, and T_{00} is the contribution of cooperation among remaining nodes. We can compute all four totals

$$T_{11} = T(\mathbf{AK}_{11}) = \sum_{a \in A_1} \mathrm{oInt}(a) = \sum_{k \in K_1} \mathrm{iInt}(k)$$

$$T_{10} = \sum_{a \in A_1} \mathrm{oExt}(a) \quad \text{and} \quad T_{01} = \sum_{k \in K_1} \mathrm{iExt}(k)$$

and finally

$$T_{00} = T(\mathbf{AK}) - T_{11} - T_{10} - T_{01}.$$

Note that we used only information from **WA**, **WK** and $\mathbf{AK}_{11}$.

Weighted degrees in derived networks

Often in a derived network, the importance of its nodes is measured by their weighted degree. It turns out that we don't need to compute the derived network to get them. They can be determined faster. Let's consider a general case

$$\mathbf{AK} = \mathbf{WA}^T \cdot \mathbf{WK}$$

where **WA** and **WK** are compatible two-mode networks. The interpretation of the network **AK** depends on the nature of networks **WA** and **WK**.

In the case when **WA** and **WK** are the authorship and the keywords matrices, the entry $ak[a, k]$ counts the number of different triples (a, w, k) such that the author a wrote the work w that is described by the keyword k – the number of times the author a is dealing with the topic k in his/her works.

Weighted degrees in derived networks

$$\mathbf{wod}_{AK} = \mathbf{WA}^T \cdot \mathbf{wod}_{WK} \qquad \text{(Oeq)}$$

$$\mathbf{wid}_{AK} = \mathbf{WK}^T \cdot \mathbf{wod}_{WA} \qquad \text{(Ieq)}$$

$$T(\mathbf{AK}) = \mathbf{wod}_{WA} \cdot \mathbf{wod}_{WK}$$

. . . Weighted degrees in derived networks

For the weighted out/in-degrees of **AK** we get

$$\mathrm{wod}_{AK}(a) = \sum_{k \in K} ak[a,k] = \sum_{k \in K} \sum_{w \in W} wa^T[a,w] \cdot wk[w,k] =$$

$$= \sum_{w \in W} wa[w,a] \cdot \sum_{k \in K} wk[w,k] = \sum_{w \in W} wa[w,a] \cdot \mathrm{wod}_{WK}(w)$$

or in a vector form

$$\mathbf{wod}_{AK} = \mathbf{WA}^T \cdot \mathbf{wod}_{WK} \qquad \text{(Oeq)}$$

and

$$\mathrm{wid}_{AK}(k) = \sum_{a \in A} ak[a,k] = \sum_{a \in A} \sum_{w \in W} wa^T[a,w] \cdot wk[w,k] =$$

$$= \sum_{w \in W} wk[w,k] \cdot \sum_{a \in A} wa[w,a] = \sum_{w \in W} wk[w,k] \cdot \mathrm{wod}_{WA}(w)$$

or in a vector form

$$\mathbf{wid}_{AK} = \mathbf{WK}^T \cdot \mathbf{wod}_{WA} \qquad \text{(Ieq)}$$

...Weighted degrees in derived networks

and finally for the network total

$$T(\mathbf{AK}) = \sum_{a\in A}\sum_{k\in K} ak[a,k] = \sum_{w\in W}\sum_{a\in A} wa[w,a]\cdot\sum_{k\in K} wk[w,k] =$$

$$= \sum_{w\in W} \mathrm{wod}_{WA}(w)\cdot \mathrm{wod}_{WK}(w) = \mathbf{wod}_{WA}\cdot \mathbf{wod}_{WK}$$

From equalities (Oeq) and (Ieq) we see that indeed both weighted degrees of **AK** can be computed faster.

Example

Cn truncations of iMetrics network and HKUST1 network

$$A_1 = \{a \in A : \mathbf{wod}_{\mathbf{Cn}}(a) \geq t\}$$

	iMetrics			*HKUST1*		
t	n	m	$avdeg$	n	m	$avdeg$
≥ 0	33919	225931	13.32	28108	45365272	3227.92
$\geq 1/10$	32418	191888	11.84	17656	3216796	364.39
$\geq 1/5$	26247	134049	10.21	10213	86529	16.94
$\geq 1/3$	14381	71967	10.01	5171	45845	17.73
$\geq 1/2$	12781	60587	9.48	4032	32806	16.27
≥ 1	6211	32395	10.43	1799	13723	15.26
≥ 2	1832	14900	16.27	689	4195	12.18
≥ 3	964	9306	19.31	369	1743	9.45
≥ 5	446	4646	20.83	172	646	7.51
≥ 10	162	1450	17.90	55	125	4.55

In the table, n is the number of nodes, m is the number arcs, and $avdeg$ is the average degree in $\mathbf{Cn}_{11}(t)$.

Truncations of iMetrics network and HKUST1 network

In the table the effects of different thresholds t on the corresponding truncated normalized co-authorship network $\mathbf{Cn}_{11}(t)$ is presented for two real-life authorship networks. It was computed using Pajek.

Note the fast decrease of the number of links m in the *HKUST1* network – the contributions of most of the authors of hyperauthored papers [6] are very small.

Conclusions

1. Although the notion of truncated networks was presented in the context of bibliographic networks the results can be applied also in other fields described by collections of networks.
2. For experimenting with smaller networks (up to 1000 nodes in each set) we developed a collection of R functions `bibmat` that supports introduced operations on networks represented by matrices. For analyzing large networks we can use their implementation in Pajek as Pajek's commands or macros.
3. We elaborated the truncated network approach for special cases such as co-appearance networks, normalized co-appearance networks, strictly normalized co-authorship networks, products of normalized networks, linking through a network, and co-citation networks. The details are available on **arXiv** [8].

Acknowledgments

The computational work reported in this paper was performed using a collection of R functions `bibmat` and the program Pajek for analysis of large networks [9]. The code and data are available at Github/Bavla [10].

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. Batagelj, V. & Cerinšek, M. On bibliographic networks. *Scientometrics* **96,** 845–864 (2013).

2. Batagelj, V. On fractional approach to analysis of linked networks. *Scientometrics* **123,** 621–633 (2020).

3. Maltseva, D. & Batagelj, V. iMetrics: the development of the discipline with many names. *Scientometrics* **125,** 313–359 (2020).

4. Matveeva, N. *Scopus data on HKUST publications in years 2017–2019. Private communication.* 2023.

5. Collaboration, C. *et al.* Combinations of single-top-quark production cross-section measurements and $|f_{\mathrm{LV}} V_{tb}|$ determinations at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS experiments. *Journal of High Energy Physics* (2019).

6. Chawla, D. S. Hyperauthorship: global projects spark surge in thousand-author papers. *Nature, NEWS* (13 December 2019).

7. Guinness. *Guinness World Record 653537: The most authors on a single peer-reviewed academic paper.* https://www.guinnessworldrecords.com/world-records/653537-most-authors-on-a-single-peer-reviewed-academic-paper. 2021.

References II

8. Batagelj, V. Weighted degrees and truncated derived bibliographic networks. *arXiv preprint arXiv:2401.04726.* Accepted in Scientometrics. (2024).

9. Nooy, W. d., Mrvar, A. & Batagelj, V. *Exploratory social network analysis with Pajek; Revised and Expanded Edition for Updated Software.* (Cambridge University Press, 2018).

10. Batagelj, V. *GitHub/Bavla/biblio bibmat.R.* https://github.com/bavla/biblio/tree/master/code. 2023.