# A Block-corrected Modularity Approach for Community Detection in Scientific Citation Networks

Hasti Narimanzadeh[1], Takayuki Hiraoka[1], and Mikko Kivelä[1]

[1]Aalto University, Espoo, Finland

Identifying communities in complex networks has long been a topic of interest and significance [1]. With academic citation networks in mind, structures within drive the mechanisms behind the creation and transfer of knowledge in and across scientific fields [2]. Many such networks come with node attributes inferred from the data which affect the structure of the networks. This thus motivates the question of how such information about the known attributes can be used to uncover mesoscopic structures, e.g., communities, that are not only explained by the known attributes but also the hidden ones.

Publication times of scientific papers induce a block structure on the network where a paper only cites papers that precede it. Recovering structures due to this temporal attribute of the scientific papers does not, however, pose much of an interesting question. Rather, a more salient question would be whether we can factor out the effect of the publication times to reveal structures that are driven by other hidden node attributes, e.g. field of research or authors' academic cooperation [3]. Many existing modularity-based community detection methods either do not address this [4], or make implicit or explicit assumptions about how blocks are connected, e.g. that the probability of an edge decays exponentially with the time difference between two nodes [5]. These assumptions are reflected in the choice of the null model.

We propose a modularity and its associated block-corrected null model for detecting communities in networks with any form of *a priori* known block structure. The probability of an edge from node $i$ to node $j$ under block-corrected null model is $P_{ij} = \frac{k_i^{\text{out}} k_j^{\text{in}}}{K_{t_i}^{\text{out}} K_{t_j}^{\text{in}}} L_{t_i t_j}$, where $t_i$ denotes the block of node $i$, $k_i^{\text{in}}$ and $k_i^{\text{out}}$ the in- and out-degree of node $i$, $K_{t_i}^{\text{in}}$ and $K_{t_i}^{\text{out}}$ the sum of in- and out-degrees of all nodes in block $t_i$, and $L_{s,t}$ the number of out-going edges from block $s$ to $t$. The block-corrected model preserves both the in- and out-degree distribution of nodes as well as the edge density between every pair of blocks. The temporal block model inherent to academic citation networks can thus be captured seamlessly through $L_{s,t}$ directly translating into the number of edges from temporal block $s$ to $t$. Here, two nodes are considered part of the same temporal block if they are published on the same day, with temporal blocks ordered by time.

Previous works have studied how attention to scientific publications decays in terms of the number of citations in citation networks, positing that an exponential temporal distribution of edges is preferred over a power law, while noting that power law decays better describe the citation patterns of more recent publication [6]. We observe a similar case when analysing citation networks of fields of Physics and Computer Science in the OpenAlex database [7] in Fig. 1. The distribution of temporal edge lengths in both of the networks seems to be better fitted by an exponential distribution for earlier publications, whereas for more recent publications, published in the past roughly 10 years (2015-2024), the distribution is more suitably fitted by a power law decay. This makes an implicit assumption of exponential decay in layer connectivity less suitable for analysing citation networks with

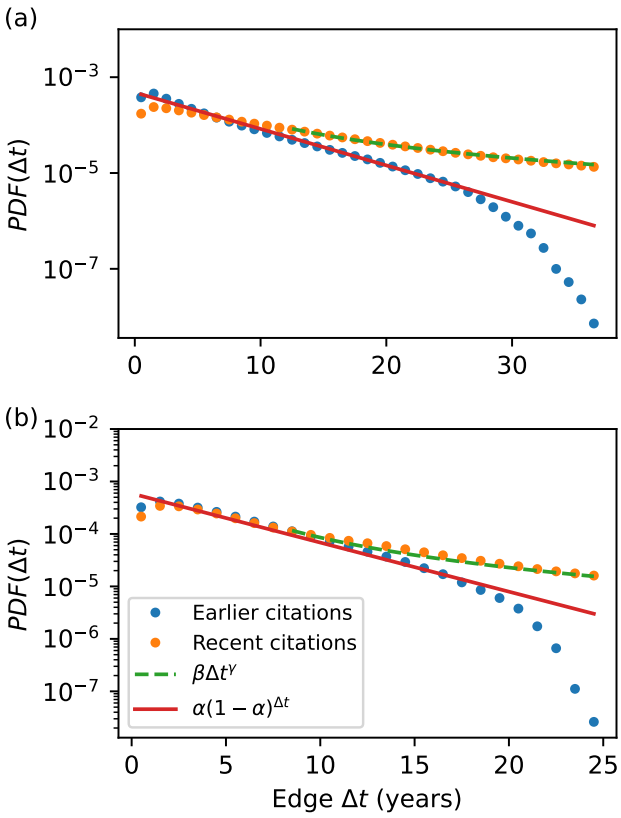


Figure 1: Distribution of temporal edge-lengths in years for citation networks of (a) Physics and Astronomy and (b) Computer Science subfields in the OpenAlex database. Earlier citations include references made by papers published before the year 2000. Recent citations include references spanning 9 years 2015–2024. An exponential is fitted to the earlier citations distributions and a power law to the tail of the recent citations distributions. Fitted curve parameters are (a) $\beta = 63.42$, $\gamma = -1.6$, $\alpha = 4.8\times10^{-4}$ and (b) $\beta = 5.0\times10^{3}$, $\gamma = -1.9$, $\alpha = 5.9 \times 10^{-4}$. Probability distribution of recent citations as a function of $\Delta t$ has a clear heavier tail compared to that of the earlier citations.

more recent data.

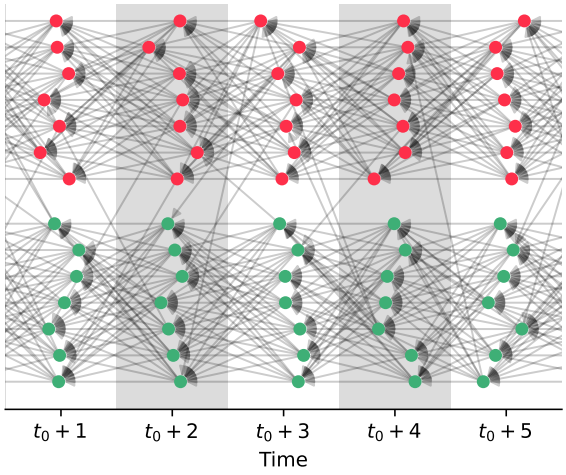


Figure 2: Schematic of a network model with two planted communities and a temporal structure where only nodes in adjacent blocks are connected. In addition, there are long links from the last block to the first.

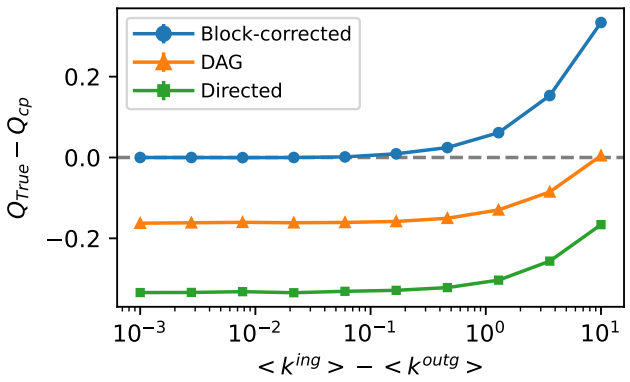


Figure 3: Difference between modularity values for the candidate partitions of the first and last ¼ of temporal blocks in one community and the rest ½ in another, and that of the true partition over an ensemble of random networks with varying average in-group degrees. All generated networks have $T_{max} = 12$ temporal blocks and $N = 100$ nodes per block. $Q_{\text{true}} > Q_{\text{cp}}$ indicates that a modularity prefers the true partition to the erroneous candidate partition.

To assess how a null model handles networks with a non-negligible number of long links, we first construct a temporal network model illustrated in Fig 2. The example synthetic network has two planted (horizontal) communities we aim to detect, and *a priori* known (vertical) temporal blocks. Nodes in temporally adjacent blocks are connected with a higher probability for those in the same community. Additionally, nodes in the first and last blocks are connected with lower probability. This synthetic network model exemplifies an exaggerated example of a heavy-tail distribution of edge lengths, with many short edges between adjacent temporal blocks, and a handful of very long ones. For each of the three modularities, we compare the modularity value of the true planted partition $Q_{\text{true}}$ to that of an erroneous candidate partitioning $Q_{\text{cp}}$, in which the first and last quarter of all nodes in the temporally ordered network are in one community and the rest in another.

For realisations of the described generative model above we compare our block-corrected modularity with those of the regular directed [4] and DAG [5]. Our results in Fig. 3 show that for this synthetic network model, the block-corrected null model proves to be impervious to the erroneous candidate partitioning ($Q_{\text{true}} > Q_{\text{cp}}$). It correctly deems the ground-truth horizontal partition preferable, even in networks whose ground-truth communities are only slightly distinguishable. The other two null models, however, erroneously prefer the incorrect partitioning over the ground-truth communities, even more so in networks whose true communities are not abundantly distinct from each other, i.e., as $\langle k^{ing} \rangle - \langle k^{outg} \rangle$ approaches zero.

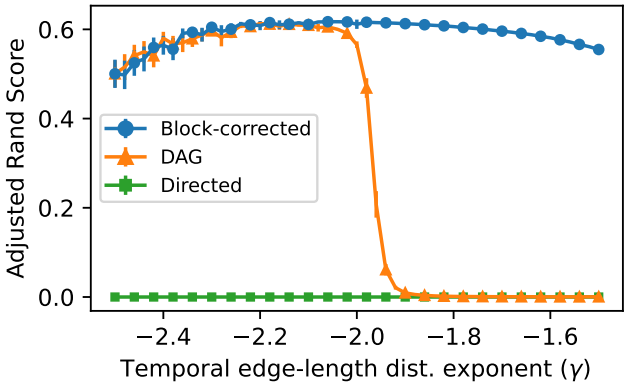


Figure 4: Adjusted Rand scores between discovered communities and the true partitions for an ensemble of synthetic networks with two planted communities. A score of 0 indicates the performance is no better than random chance. The synthetic networks are generated with a fixed number of nodes ($N = 200$) per temporal block ($T_{max} = 200$). Each node has an in-group degree $\langle k^{ing} \rangle = 8$ and out-group degree $\langle k^{outg} \rangle = 4$. Edges probabilities decrease as $P(\Delta t) \propto \Delta t^{\gamma}$. All three models are coded to find exactly two communities. Error bars show 95% confidence interval. Our proposed block-corrected method (blue) remains consistent across the range of exponents while the performance of the DAG null model (orange) sharply drops below a certain threshold for $\gamma$. The regular directed modularity (green) fails to capture the planted community structure.

A synthetic network model where the temporal edge lengths follow a smoother probability distribution embodies a more realistic setting, similar to one observed in real-world data in Fig. 1. For this reason, we construct a second generative network model to understand the behaviour of the three null models. This generative model is a random directed acyclic graph (DAG) model with time stamps where the temporal edge lengths distribution resembles a heavy-tailed power law. We then draw an ensemble of such temporal networks with two sufficiently distinguishable planted communities and a fixed number of temporal blocks. The quality of a partition is quantified by the adjusted Rand score between that and the true partition. A score of 0 represents performance no better than random labelling of nodes and 1 is a complete match between the candidate partitioning and the true one. Figure 4 illustrates as the edge length distri-

bution tail becomes heavier, that is, as the magnitude of the exponents becomes smaller, the block-corrected retains its ability to distinguish the ground-truth communities, whereas a sharp drop is observed in the case of the DAG modularity. The directed regular modularity fails to capture the temporal citation patterns in the networks altogether, and its adjusted Rand score remains at zero across all exponents.

Publications in the OpenAlex database are accompanied by their scientific fields, with different levels of granularity, ranging from topic and sub-field to scientific domain, as well as information about publication venues, author affiliations, and funding sources. Such abundant and varied types of metadata provide an opportunity to study whether temporally invariant communities induced by them can be recovered, discounting the known temporal patterns in the citation networks.

**Conclusion.** Observations from scientific citation datasets signify a change in citing pattern moving from an exponential distribution of time differences of citation edges to a more heavy-tailed distribution. This necessitates a re-evaluation of community detection methods for their usability in the domain of citation networks. In this work, we studied the shortcomings of some of the existing modularity-based approaches when faced with heavy-tailed synthetic citation networks. We then proposed a block-oriented modularity, which is more generalisable to arbitrary distributions of temporal connectivity without compromising scalability or computational efficiency. This opens up an avenue for researchers toward a more reliable use of modularity-based community detection methods on more recent citation network datasets.

# References


[1] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[2] Derek J De Solla Price. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515, 1965.

[3] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[4] Elizabeth A Leicht and Mark EJ Newman. Community structure in directed networks. *Physical review letters*, 100(11):118703, 2008.

[5] Leo Speidel, Taro Takaguchi, and Naoki Masuda. Community detection in directed acyclic graphs. *The European Physical Journal B*, 88:1–10, 2015.

[6] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A Huberman, Kimmo Kaski, and Santo Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734–745, 2015.

[7] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv:2205.01833*, 2022.