Visualization of the Vispubdata [5]

Avantika Mudumbai

IMT2019013

Abstract

Vispubdata.org is a metadata collection of every paper that has appeared at the IEEE Visualization (VIS) set of conferences: InfoVis, SciVis, VAST, and Vis. It contains attributes like title, abstract, authors, and citations to other papers in the conference series for each of the publication. In this project, we aim to understand the dataset better by creating visualizations of aspects like temporal similarity between authors.

This is the link to the github repository containing the code used

1 Problem Statement

We use the vispubdata and create 3 network layers:

1. **Co-authorship layer:**
$Graph : G(V, E)$
$V : \{u : u \in Authors\}$
$E : \{(u, v) : u, v \in V$ such that there is at least one paper where u and v are co-authors$\}$

2. **Author-Topic similarity layer:**
$Graph : G(V, E)$
$V : \{u : u \in Authors\}$
$E : \{(u, v, w) : u, v \in V, w = similarity(u, v)\}$

3. **Temporal similarity layer:**
$Graph : G(V, E)$
$V : \{u : u \in Authors\}$
$E : \{(u, v, w) : u, v \in V, w =$ distance between the time series data of frequency of publications of u and $v\}$

We then separate the nodes, ie the authors, into communities and then analyse trends within as well as across communities.

2 Data

The dataset includes metadata on papers that appeared at the IEEE VIS conference series from 1990–2021. For each paper, there are the following fields:

- **Conference:** Name of the conference in which the paper appeared in (InfoVis, SciVis, VAST, or Vis). See figure 1 for the statistics.

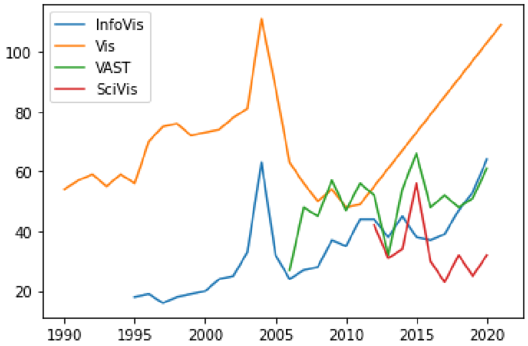
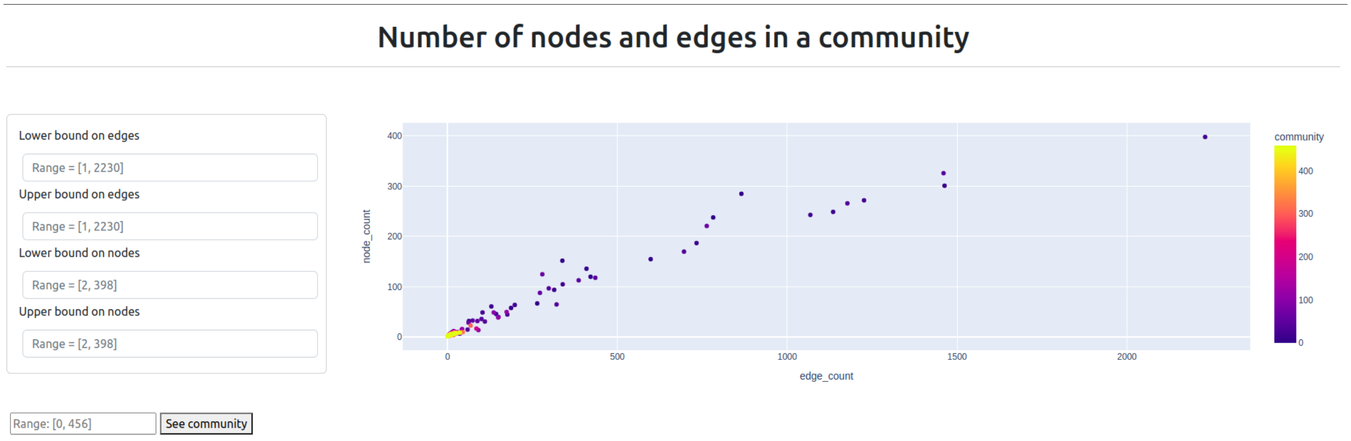


Figure 1: Number of publications for a conference

- **Year:** Year in which the paper was presented in the conference. (Note: This is not necessarily same as year of publication)
- **Title:** Title of the paper
- **DOI:** The paper's DOI pointing to a digital library entry.
- **Link:** The link to the paper in the IEEE digital library.
- **FirstPage:** The number of the paper's first page in the printed proceedings or the journal special issue
- **LastPage:** The number of the paper's last page in the printed proceedings or the journal special issue.
- **PaperType:** : one of C (conference paper), J (journal paper), M (miscellaneous: capstone, keynote, panel, or poster)
- **Abstract** as given in the publication.
- **AuthorNames:** Names of the authors of the publication separated by a semi-colon.
- **AuthorAffiliation:** The organizational affiliation of the first author
- **InternalReferences:** A list of references this article makes to other IEEE VIS papers, using the unique identifiers of the DOI column. We



Number of nodes and edges in a community

Lower bound on edges

Range = [1, 2230]

Upper bound on edges

Range = [1, 2230]

Lower bound on nodes

Range = [2, 398]

Upper bound on nodes

Range = [2, 398]

Range: [0, 456]

Figure 2: Home Page of the dashboard

do not include citations to papers outside of the conference series.

- **AuthorKeywords:** A list of author keywords supplied on the paper PDF

3 Pre-Processing

3.1 Co-Authorship Layer

Louvian community detection is a method to extract communities from large networks created by Blondel et al. [4]. The method is a greedy optimization method that appears to run in time $O(n \log n)$ where n is the number of nodes in the network.

This layer is an unweighted graph where an edge between 2 nodes(authors) denotes that the 2 authors have co-authored at least one paper together. We use the Louvian community detection algorithm on this graph using an existing implementation in the python networkx library [1]. The algorithm separates the network into 460 communities. The file *dataset/layer1/community_stats.csv* contains the information about number of nodes (authors) in each community.

3.2 Author-Topic Similarity Layer

Word2vec [6] is a technique for natural language processing (NLP) that uses a neural network model to learn word associations from a large corpus of text. Each distinct word is represented using a vector of numbers. The vector is chosen such that even a simple function like the cosine similarity can indicate the semantic similarity between the words that are represented by those vectors.

The second layer of the network is a weighted graph which contains all the edges in the first layer. The weights of the edges are determined using the similarity function with the *en_core_web_md* model from the SpaCy library [2] in python. This similarity function uses the combination of the cosine function with the word2vec algorithm. The vector corresponding to each node(author) consists of the names and index terms of their publication(s). This results in higher similarity scores for the authors who have co-authored multiple papers than those who have co-authored only one

3.3 Temporal Similarity Layer

The Pearson correlation coefficient [7] is a statistical measure that gives a linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations which is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1. Here, 1 signifies perfect correlation (linear dependence), 0 signifies no correlation and -1 signifies perfect negative correlation(inversely dependent).

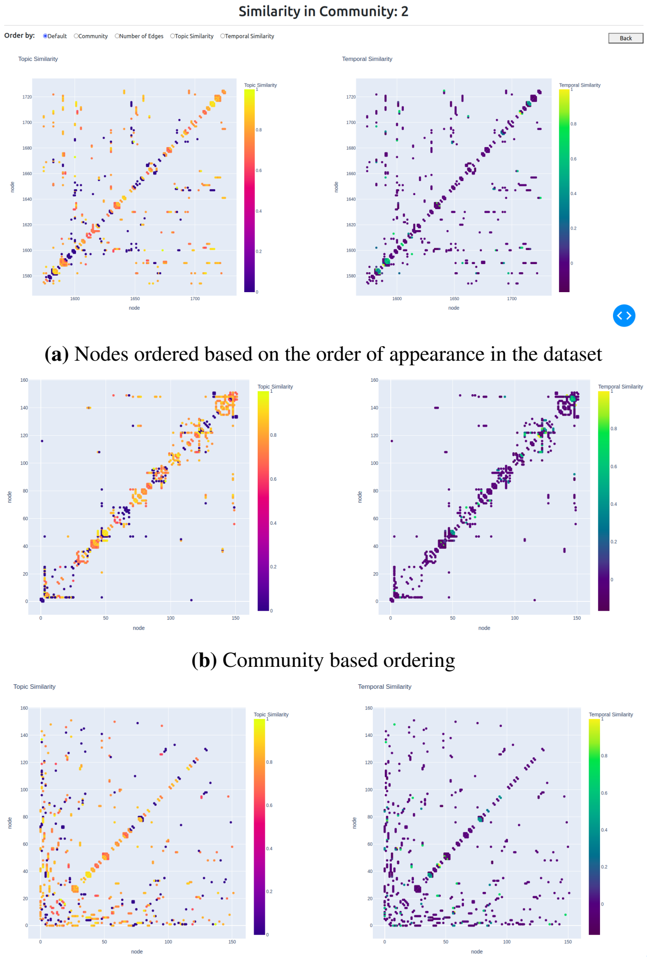
This layer is a weighted graph which contains all the edges in the first layer. The weights of the edges are determined using the pearson correlation of te arrays of the number of their publications in an year. This array consists of 32 elements mapping to the years [1990, 2021] (both inclusive). The Pearson co-efficient is calculated using the *corrcoef* function in the *numpy* library.

4 Dashboard and Visualizations

The dashboard consists of 2 pages: home page and the simiarity page. The homepage, as seen in figure 2, consists of a graph that denotes the number of nodes and edges in a community. One can also give a custom range of number of nodes and/or edges, and only the communities that satisfy the range will be shown on the graph. We can use the see community option to see the topic and temporal similarity of authors in the community given as input.

The similarity page, as seen in figure 3, shows the topic and temporal similaritites between the authors in the community 2. There are options to seriate the matrix based the following:

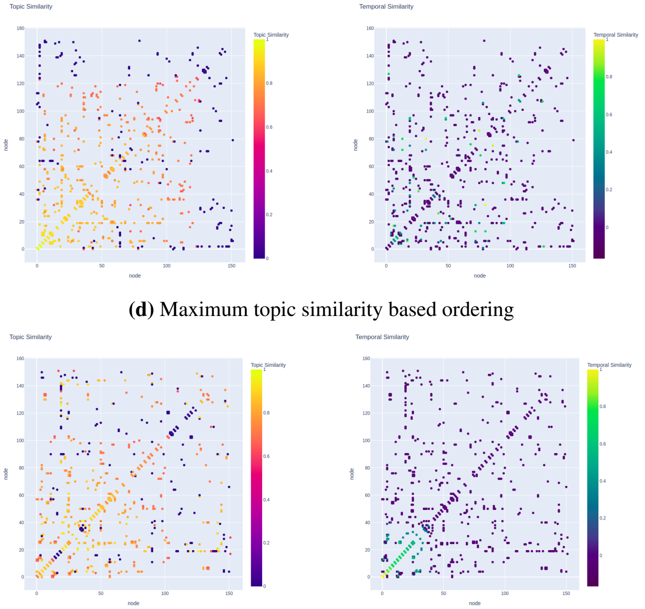
- **Community:** Based on communities detected within the community.
- **Number of Edges:** The nodes are ordered based on the number of edges.
- **Topic Similarity:** The nodes are ordered by the maximum of their topic similarity with the other nodes.
- **Temporal Similarity:** The nodes are ordered by the maximum of their temporal similarity with the other nodes.


Similarity in Community: 2
Order by: ● Default ○ Community ○ Number of Edges ○ Topic Similarity ○ Temporal Similarity
Back
Topic Similarity
Temporal Similarity
node
node

(a) Nodes ordered based on the order of appearance in the dataset

(b) Community based ordering

(c) Nodes ordered based on number of edges

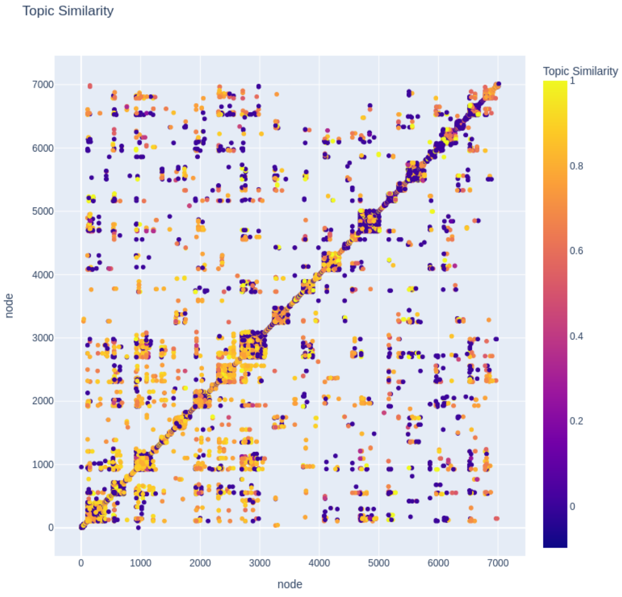


(d) Maximum topic similarity based ordering

(e) Maximum topic similarity based ordering

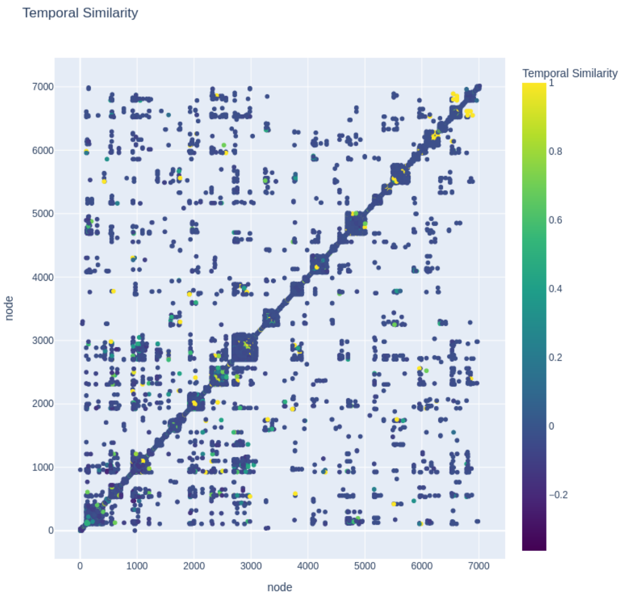
Figure 3: Matrix Seriation in community 2

5 Observations and Inferences



(a) Layer 2: Topic Similarity in the Network

- Figure 4 depicts the author topic and temporal similarity in the entire network. The nodes have been re-ordered according to the community. As seen in figure 4, we observe that the authors have co-authored with almost all authors in their community. We also observe that there is high author-topic similarity especially

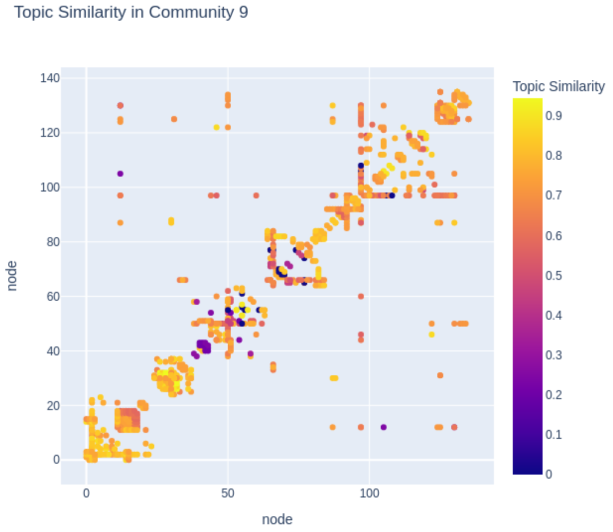


(b) Layer 3: Temporal Similarity in the Network

Figure 4: Topic and Temporal similarity between authors in the network

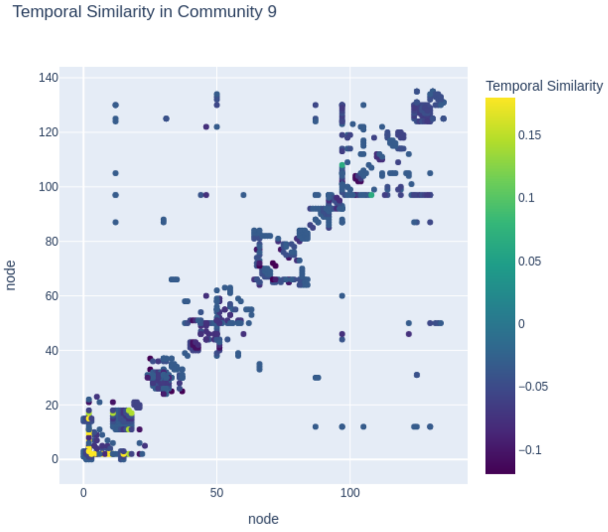
among those in the same community. But, we also oserve that there is low temporal similarity amongst the authors.

- In community 2, see figure 3b, we see that even within the community, the nodes can further be divided into communities, such that the authors have co-authored with other authors from the same sub-community (with a few exceptions).



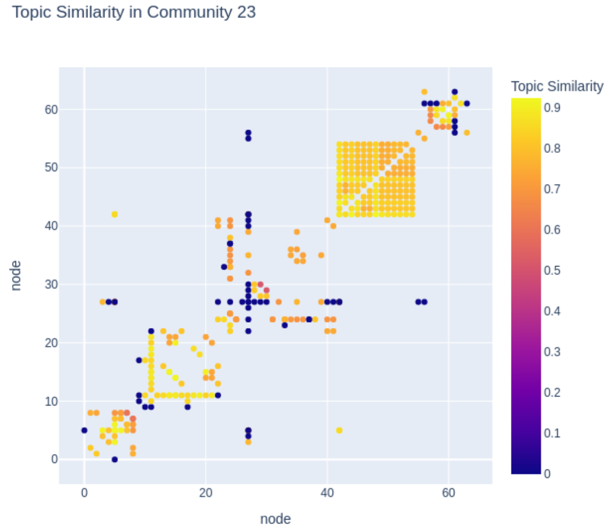
(a) Topic Similarity

- In community 9, we observe that there is extremely high author topic similarity, but low temporal similarity. From this, we can infer that while the authors have similar topics of research, they have not co-authored multiple papers together (due to low temporal similarity).



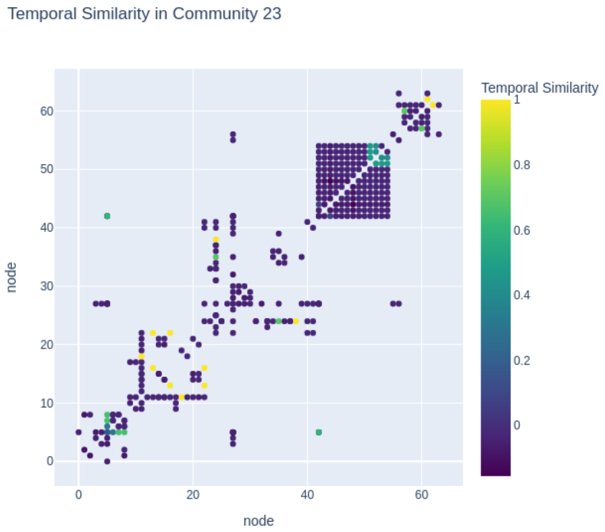
(b) Temporal Similarity

Figure 5: Similarity in community 9



(a) Topic Similarity

- In community 23, figure 6, we see that there is a group of about 60 authors who have worked with each other and have similar topics of research. But due to their negative temporal

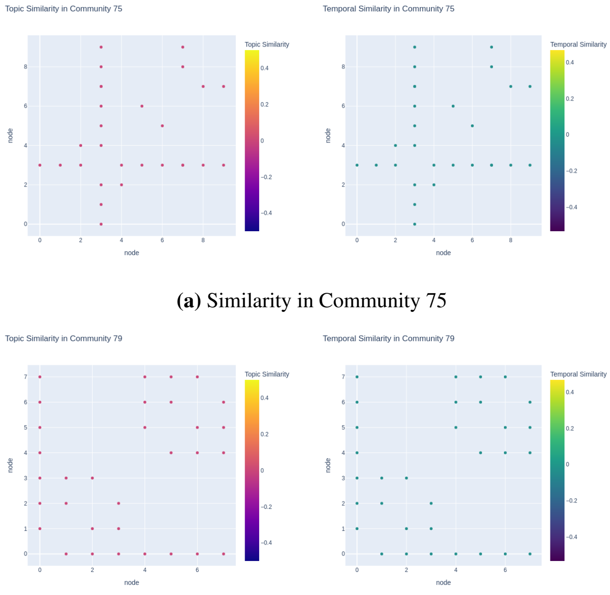


(b) Temporal Similarity

Figure 6: Similarity in community 23

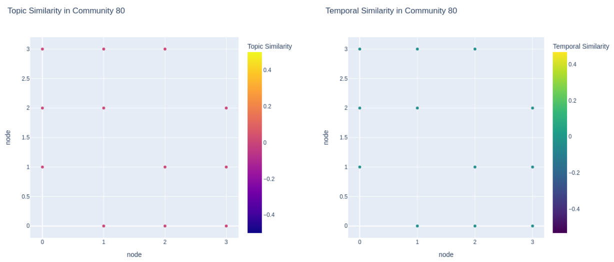
similarity, we can conclude that they have not worked on multiple papers together. This is with the exception of those few pairs of authors who have positive temporal similarity (≈0.5).

- We also see that there are more about 40 communities that have no topic or temporal similarities. Out of these communities, about 35 of them contain only 2 nodes.



(a) Similarity in Community 75

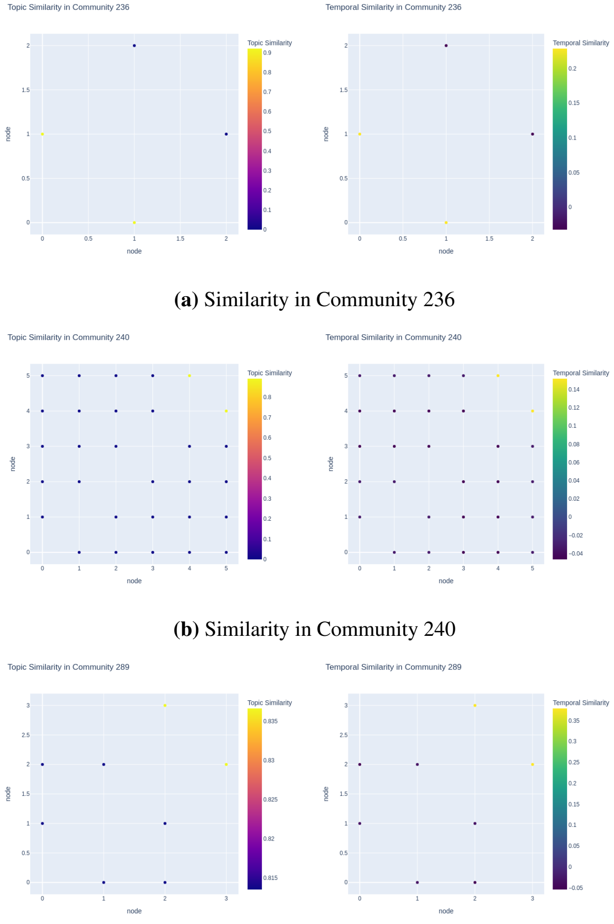
(b) Similarity in Community 79



(c) Similarity in Community 80

Figure 7: Few communities with more than 2 nodes and no similarity

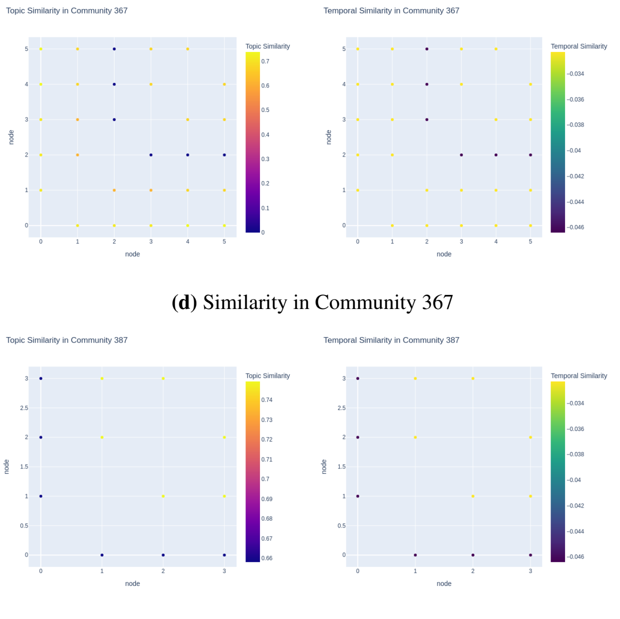
- There are few communities, see figure 8, whose topic and temporal similarities are proportional. We observe that these communities have low or negative temporal similarities and low topic similarities. From this we can conclude, while the topic and temporal similarities are linearly co-related, the authors have not co-authored multiple publications and have mostly different topics of research.



(a) Similarity in Community 236

(b) Similarity in Community 240

(c) Similarity in Community 289

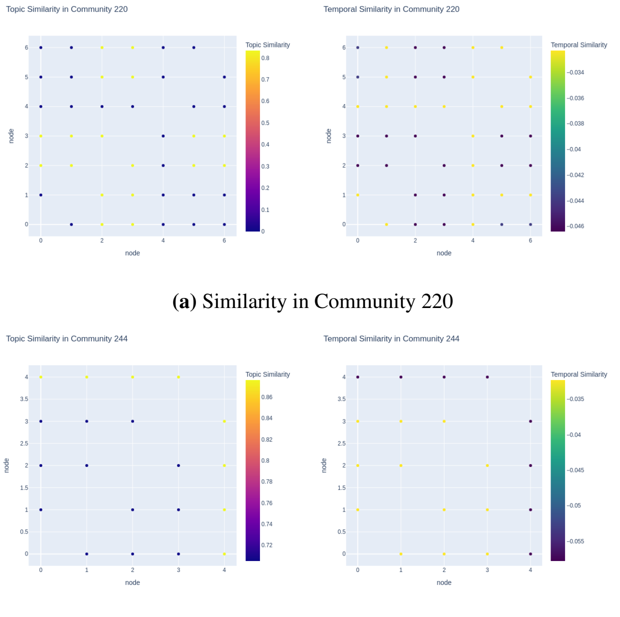


(d) Similarity in Community 367

(e) Similarity in Community 387

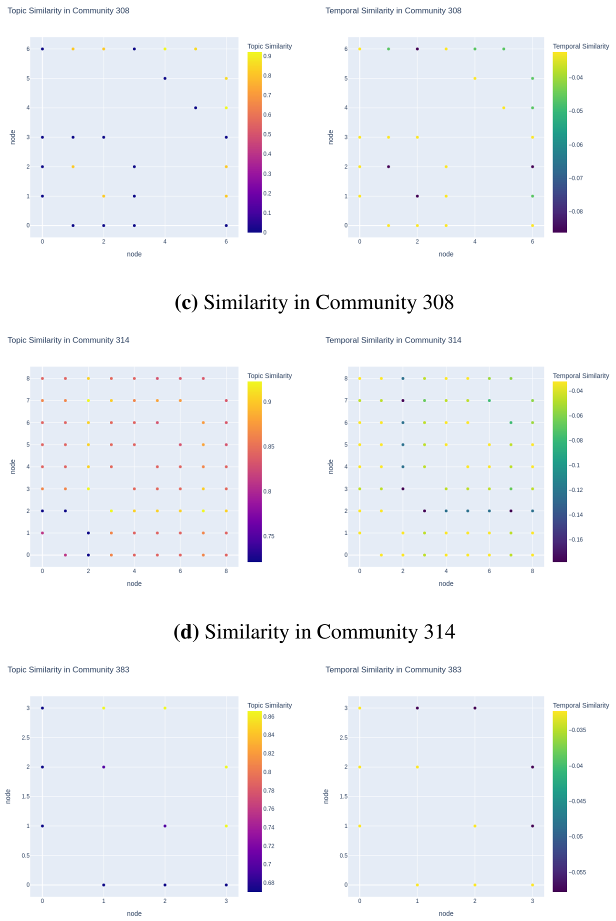
Figure 8: Communitites with proportional similarities

- There are few communities whose topic and temporal similarities are inversely proportional. We also observe that the topic similarities are high where as the temporal similarities are mostly negative. From this, we can conclude that the authors have similar areas of research but have not co-authored multiple papers.



(a) Similarity in Community 220

(b) Similarity in Community 244



(c) Similarity in Community 308

(d) Similarity in Community 314

(e) Similarity in Community 383

Figure 9: Communities whose similarities are inversely proportional

- Out of the 465 communities, there are more than 400 communities that have less than 10 nodes (authors). Even amongst these communities, there are 139 communities that consist of only 2 nodes. Since we do not look into inter-community edges, the trends that could be set by these nodes are missed.

6 Conclusion

While the temporal similarity, ie the similarity between frequency of publications, is low among authors, the topic similarity is relatively higher. The communities that have been formed mostly consist of authors who have similar topics of interest. We can also conclude that these authors have not co-authored multiple papers with the same author as even when the topic similarity is high, the temporal similarity is low.

References

[1] Networkx documentation.

[2] Spacy documentation.

[3] S. Bani-Ahmad, A. Cakmak, G. Ozsoyoglu, and A. Al-Hamdani. Evaluating publication similarity measures. *IEEE Data Eng. Bull.*, 28:21–28, 01 2005.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[5] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Vispubdata.org: A metadata collection about ieee visualization (vis) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, 2017.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[7] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1901.

[8] J. Sawatphol, N. Chaiwong, C. Udomcharoenchaikit, and S. Nutanong. Topic-regularized authorship representation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[9] C. R. Sugimoto, C. Ni, J. D. West, and V. Larivière. Great minds think alike, or do they often differ? research topic overlap and the formation of scientific teams. *Research Policy*, 42(3):765–776, 2013.

[10] D. C. Zhang and H. W. Lauw. Variational graph author topic modeling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’22, page 2429–2438, New York, NY, USA, 2022. Association for Computing Machinery.