# PERFORMANCE ANALYSIS OF THE COMMUNITY RECONSTRUCTION TECHNIQUE USING A HOMOGENEOUS COLLABORATION MULTIPLEX

-By Dhanyashree Yelahanka Srinivasa and Sudeepkumar Shetty

**Introduction:**

Finding communities and patterns in multiplexes has several uses in different domains. In this project we find communities in the multiplex and try to analyse them and find interesting patterns.

**Dataset description:**

The Dataset we used for this project is the DBLP dataset. The DBLP Computer Science Bibliography contains more than 3.66 million journal articles, conference papers, and other publications on computer science (as of July 2016). All important journals on computer science are tracked by DBLP.



*Fig 1: snapshot of the dataset*

The figure one shows a snapshot of the raw dataset which is in Xml format.

**Approach:**

We first obtained the data set and then pre-processed the data. The pre-processing phase includes the following steps:

* We initially converted the file from xml format to excel and dropped a few unnecessary columns
* We then further reduced the size of the dataset by considering data only since 2003 that is in other words we considered 15 year’s worth of data
* This reduced data contained journals, articles and conferences. But for this project we only interested in journals so we filtered out all the journals.
* Building the multiplex:
  + We then analysed the data and decided upon the 5 journals we wanted to use to build the layers of the multiplex.
  + The five journals we choose were
    - IEEE Transactions on Computers
    - ACM Transactions on Computer Systems
    - Future Generation Comp. Syst.
    - IEEE Computer
    - International Journal of Computer Vision
  + Our reason for choosing these journals was we wanted to analyse some interesting patterns so we thought of picking 4 journals from the same field and one journal from a different field. We also wanted to considered journals with different impact factors to see if impact factor had any impact on the communities.
  + So we choose 2 journals with fairly high impact factor from the same field. Namely:
    - ACM Transactions on Computer Systems – 5.045
    - Future Generation Comp. Syst. – 3.977
  + So we choose 2 journals with comparatively low impact factor from the same field. Namely:
    - IEEE Transactions on Computers-2.916
    - IEEE Computer - 6783 -impact factor - 1.755
  + We then choose one journal with high impact factor but from a different field, namely
    - International Journal of Computer Vision - 8.222
  + We then started building the multiplex.
  + We considered all the authors who had published in any of these five journals to be the nodes of the multiplex and if these authors have worked together then they have an edge between them.
  + Once we had built the multiplex we then used the Infomap code to detect communities in multiplex.
  + We then analysed the results and tried to determine interesting patterns. We looked at each layer based on the number of communities, size of communities and so on.

**Analysis and Observations:**

We came up with 5 questions to analyse the multiplex based on the communities.

1. We first wanted to see if there is any relationship between the impact factor and the number of communities in the individual layers.

Fig 2: a bar graph indicating the number of communities in each layer

We can observe from fig 2 and fig 3 that the journal with the highest impact factor is the IJCV journal with an impact factor of 8.22 but we can also observe that the ATCS journal with an impact factor of 5.045 has the most number of communities. But when consider just the journals in the same field we still see the same pattern and hence we conclude that impact factor doesn’t really have impact on the number of communities.

Fig 3: the impact factor of all the journals.

1. We were curious to know if a group of authors working together in a particular domain also worked together in a different domain.

Fig 4: collaborations across domains

As we can observe from fig 4 the IEEE Computer journal had the highest number of common collaborations with the International Journal of Computer Vision which is from a different domain.

We can also notice from fig 4 that 2 high impact factor journals that we considered have no common collaborations. We can infer from this fig that as the impact factor decreases the number of common collaborations across domains increases.

This also answers our **question that if any author publishing in a high impact factor journal in a particular field will also publish in a high impact factor journal in a different field?**  we can assume from our results that there are no common authors between the two high impact factor journals we choose for our project and no author might publish in 2 high impact factor journals of 2 different fields.

1. We then wanted to know if a group of authors publishing in high impact factor journals also collaborate with each other in low impact factor journals.

Fig 5: the series one indicates the number of communities and series 2 indicates the size of the largest community.

From this graph we can infer that some authors collaborating with each other in an high impact factor journal also collaborate with each other in low impact factor journals.

1. We then wanted to find the best communities in each journal.

we considered a community to be the best community if it had the highest score. We computed the score by taking the product of the community density and the number of vertices.

|  |  |  |  |
| --- | --- | --- | --- |
| Journal | #of vertices | Community density | score |
| FGCS | 1480 | 0.00711 | 10.522 |
| ATCS | 31 | .76 | 23.66 |
| IEEEC | 41 | .496 | 20.34 |
| IJVC | 134 | .077 | 10.33 |
|  | 48 | .278 | 13.36 |

1. Nodes with highest involvement in and composed layers.



Here, we have a list of 10 nodes with highest involvement across all our AND composed layers. ITC looks like the most favoured journal from this analysis, IJCV is the least favoured journal. The fact that it is from a different field becomes evident here. When any 2 journals are considered, we come to know that if the author is present in the AND composed layer of those journals or not.