Application of bibliographic data-analysis and clustering for policy makers applied to hypersonic terms in Australia

Raoul Mazumdar

Melbourne, Australia

Abstract

Data analysis and primarily, clustering, of bibliographic information can yield information for policymakers in technology canvassing, current research clusters, and provide a convenient interface for data interrogation. The use of clustering metrics can show researchers whom the most influential members or institutions are within a specific technology area, thereby assisting in the tendering process of any prospective technology and policy. Furthermore, the information can unveil the current nature of the research work in the domestic domain, and help policies focus on potential opportunities. The approach has been applied to general hypersonic related terms, including scramjets and ramjets in this work. The outcome shows that the University of Queensland and the University of New South Wales at Canberra are influential partners in this domain. This is not necessarily new information, as this is well depicted in the abundance of papers present in hypersonic literature. However, the clustering metrics highlight other influential authors within other governmental agencies and the University of Southern Queensland and the Royal Melbourne Institute of Technology. The clustering work also highlights several different private organisations who have contributed in some form to the Hi-Fire hypersonics research program, further emphasising the potential of clustering as a tool for policy development. The cluster's general structure and interaction and the applied nature of hypersonic research indicate that domestic collaboration is underutilised and could be enhanced if a policy is developed to address the unconnected research groups.

Keywords: Clustering, Bibliographic, Data-analysis, Hypersonics, Policy

1. Introduction

From a policy point of view, it is crucial to know the current research level and collaboration within the prospective technological area to address them better. If a field is shown to lack academic research and staff, the policy can be driven to develop the fundamental research. Whereas if a particular technology area shows a large degree of research and development, the policy can help technology diffusion towards the application phase. Understanding the interconnectivity of a research cluster, key staff members, and growth is valuable. In semiconductors, quantum computing and other niche areas, canvasing would often require a subject matter expert to disseminate the information. On the other hand, clustering and bibliographic analysis can be re-purposed as a fast canvasing tool and visualisation aid, for the inspection of technologies, their clusters, and critical members for engagement.

Clustering of publications is not a new activity in itself, and have been applied to several patent databases [1, 2]. In other instances, publications have been analysed concerning geography as per the works of Baptista [3]. The paper shows how proximity can be strongly correlated with enhanced technology diffusion and innovation. Additional research by Liyanage [4] further reinforces a positive coupling between research collaboration and innovation clusters. The bodywork suggests that co-authorship of journal publications are linked to positive research outcomes. The research clustering's volume or size in different institutions also holds weight in terms of publication, productivity, and general output. Dundar and Lewis [5] discussed this, investigating productivity in universities within the United States across several disciplines. The findings show again, correlation to the size of a department and group and their research output, including journal publications. The strength of the correlation happens to be specific to the scientific field. Fundamental research activities are shown to garner more international collaboration. The paper also expresses an issue where domestic entities are less likely to collaborate due to competitive advantage [6]. Unfortunately, this lack of domestic collaboration is more pronounced with application-oriented research, in which mature hypersonic research would coincide within.

It is important to note that co-authorship does not necessarily indicate cross-collaboration, visiting PhD students, research exchanges, and one-time projects can produce those co-authorships without long term cooperation re-

search. However, the pattern of the co-authorship clusters can shed light as to the nature of the relationship. The works of Velden et al. [7] describe the reasoning behind the particular formation of clusters, indicating whether clusters are a result of cooperative agreements, staff migration or dominated by a single principal investigator. This leads to a theoretical study of clustering by Molontay and Nagy [8]. The work shows a strong correlation between individuals with high centrality in the cluster, and high citation counts and other scientometric values. Co-authorship is also shown to be weakly coupled to the citation, thereby higher co-authorship is likely to produce publications with more significant citations. Articles also receive higher citations when they expand upon a group of publications, thereby closely tied to multiple sources and hence the possibility of further citations[9]. Thus individuals with high centrality in bibliographic clusters are likely to contribute to large research outputs, regardless of the journal field.

Figure 1 a.) shows a co-authorship cluster centred around a primary node. This node generally has an abundance of co-authorships and is generally the principal investigator institution or university. The second case in Fig. 1 b.) highlights the connection of a single node between two main clusters, the reasoning for this can stem from staff migration, visiting researchers, and or one-off funded projects. Generally speaking, they indicate temporary positions and academic staff's movement, thus leading to the co-authoring pattern as presented. In Fig 1 c.) we can visualise what is likely to be an international cooperative research initiative or some domestic research hub. The multiple large nodes represent small interconnected research groups, and can be considered multiple principal investigators closely linked together. Multiple connections between separate nodes with the principal investigator represent large cooperative agreements, broad research exchanges, joint-PhD students and or the application of complementary knowledge.

Following this, the work looks to investigate research clusters applied to an area of hypersonics, which includes the re-entry of space vehicles, and high-speed aircraft and rockets. The area has undergone intensive development since the 1960s but remains a resource-intensive field [10]. First, the flow-speeds and high enthalpies can only be recreated in shock tubes and shock-reflection tunnels, which generate milliseconds of applicable air-flow. Furthermore, the computational tools coupled with reacting chemistry is numerical burdensome. These factors, coupled with fundamental research and

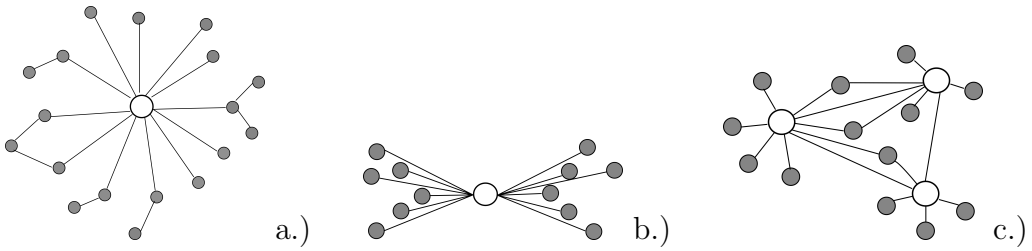


Figure 1: Typical clustering arrangements adapted from Velden et al. [7]

laser diagnostics' growth, have slowly allowed the maturity of the technological field. That said, within Australia, the development of a mature hypersonic system is still pending. Thus, examining the research cluster within Australia is an area of interest, and how academic and industry partners are collaborating through journal publications.

In this body of work, bibliographic journal publications about hypersonics are analysed using general data analytics and a clustering-based method. Due to the field's sensitivity, journal papers are only sourced from keyword searches that include hypersonics, scramjets or ramjets. This happens to exclude other fundamental research areas and complementary technological areas. The research is conducted through the use of data-harvesting from the Scopus journal database and followed by post-processing through python based programs to sort names and university affiliation. The processed information is then imported into Gephi1, open-source software for manipulating and exploring networks as depicted by Fig. [11].

2. Method

2.1. Data Harvesting and Sorting

Various journal publications are sourced through Scopus using an application programming interface written in Python for importing journal authors, affiliations, abstracts, keywords and publication dates. However, as mentioned in papers concerning bibliographic data-analysis, name identification is a problem [7]. Different journal publications hold different name conventions and express a single author in different formats, thereby alluding to multiple authors instead of one. In order to account for this, each new name in a data-set is checked for similarity for any other names within the data-set.

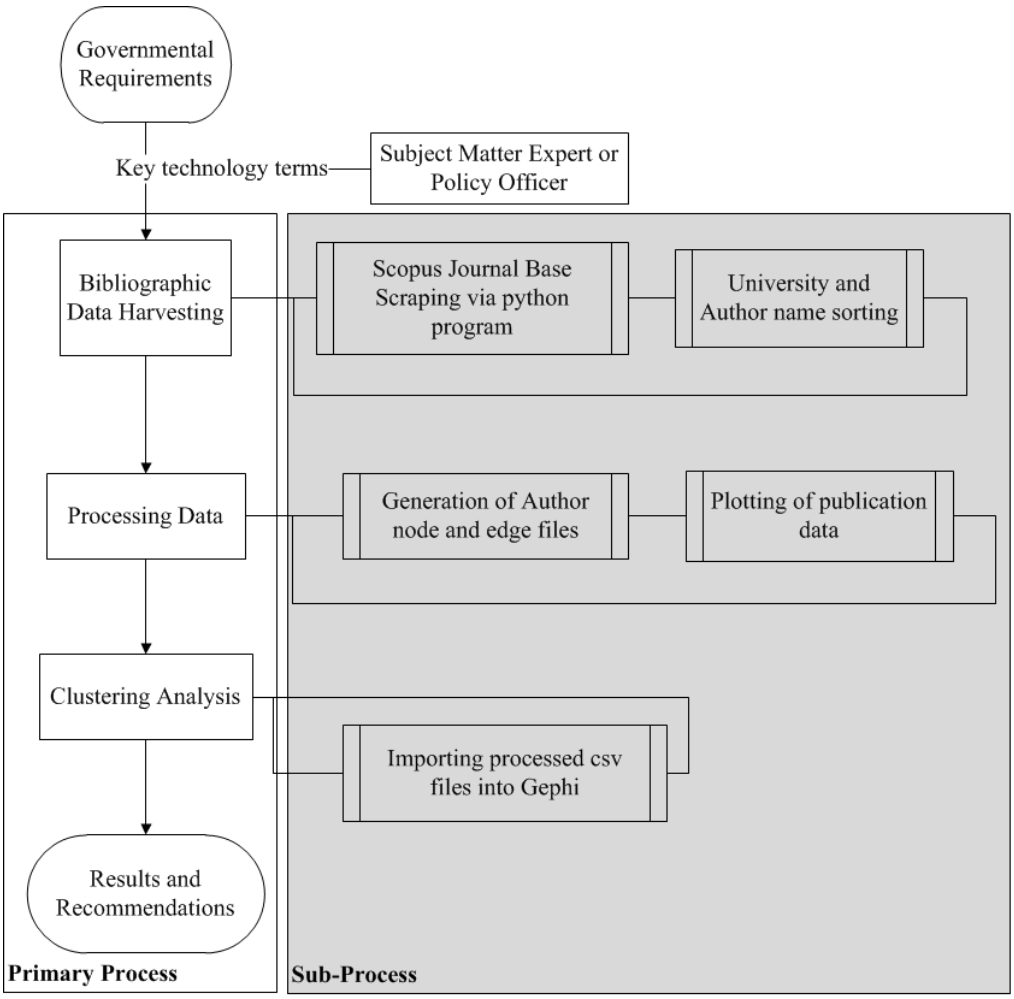


Figure 2: Data-analysis and method

The process works by generating a list of possible permutations of a name and checks the prospective strings using Levinstein distance [12].

Similarly, domestic universities are also subject to different naming conventions. For example, the University of New South Wales can also be represented as UNSW Sydney. A solution includes the checking names against a common set of aliases within the python program using the Levinstein distance. Furthermore, prefixes and suffixes are added and checked for a high degree of similarity from a target string. These words include common terms that may be used, such as, school of engineering or engineering department.

While these approaches are effective for most data strings, there are errors due to the tolerance required for name similarities to be met.

2.2. Clustering and Mapping

Upon completing the previous task of sorting the data into two formats, nodes and edges. In which nodes represent an author affiliation, and edges represent an author's list of collaborators. The material is ready to be imported into an open-source software called Gephi. This software allows for the visualisation of nodes in clustered maps, accommodating various clustering modes. Modes such as force-atlas, where nodes act as point charges but edges act as attractive masses. Thereby leading to well-connected nodes forming at the heart of a group and disconnect authors pushed to the exterior. In this work, a Fruchterman Reingold approach is used, which provides better visualisation for independent groups. Although the clustering map is more comfortable to visualise, there can be thousands of nodes in some cases. While examining all the nodes might be fruitful, the time required is impractical. Through Gephi, it is possible to conduct statistical analysis for nodes that are centrally position. The term, betweenness centrality, represents the shortest distance between other co-authors. They can represent principal investigators, mobile staff, or generally well connected or influential individuals. Using this metric, we generate a list of well-positioned individuals which researcher can instantly engage on for policy or tendering processes.

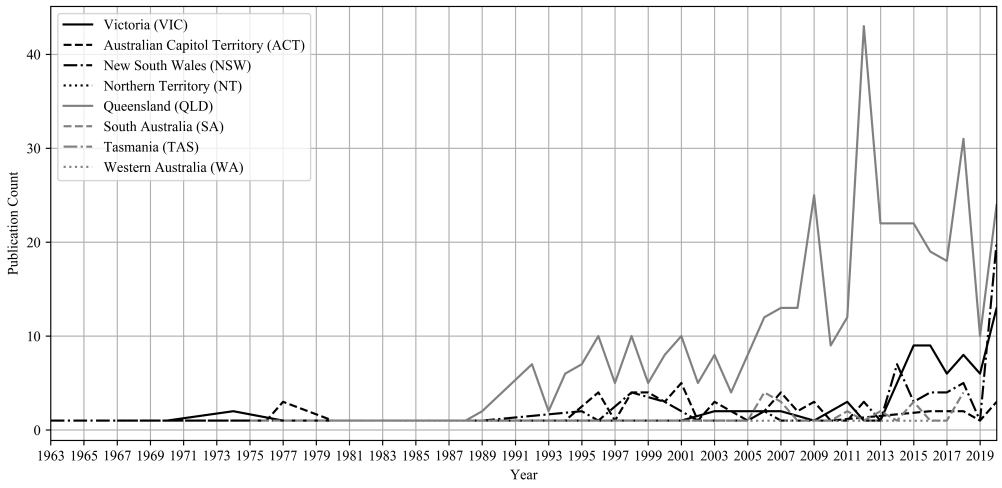
2.3. Result

The bibliographic information's initial findings show a large publication volume from the state of Queensland, followed in part by New South Wales, and Victoria as per 2.3 a.). The Australian Capital Territory is highly influential in having an early publication history dating back to the 70s, but is under-represented in the publication count. The Australian National University and the University of New South Wales at Canberra have been active in hypersonic related research. However, the University of New South Wales at Canberra is a satellite campus of the University New South Wales at Sydney due to the data-analysis and arrangement. Hence, publications from that satellite campus are merged with the primary university, thus overestimating the contribution of New South Wales while under-representing the Australian Capital territory. That aside, we can also see the country affiliation of co-authors in the data-set in Fig. 2.3 b.). The entire volume of

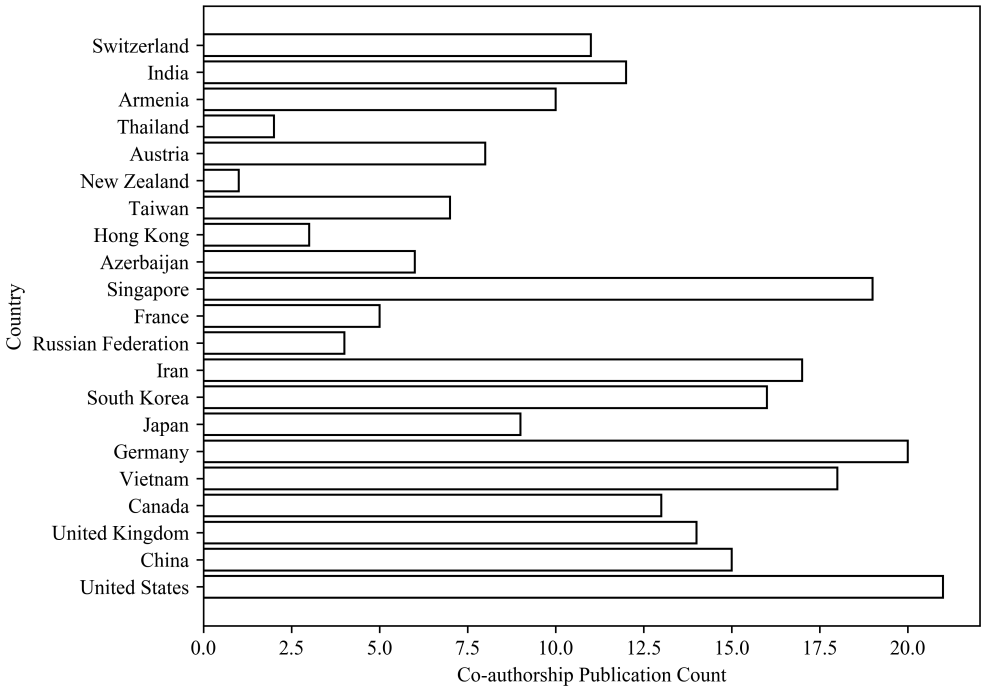
overseas authors is relatively small in the data-set, compared to the volume of papers published. Interestingly the largest volume of foreign co-authorship is from the United States, Germany and Singapore in that order. Examining the geography information can provide us with insight into the general research trends, but this excludes various other information. If we wish to understand the key players in the data-set, a clustering approach is further required.

A more visually informative approach is shown by clustering the data using the open-source software Gephi and the imported processed datasets. Figure 2.3 shows the full co-authorship cluster, which holds 972 authors represented as nodes, including 4990 edges representing co-authorship of papers. Around 42% of the nodes represent either an international entity or a domestic industry as a co-author. The rest of the nodes are derived from some author within an Australian university. The second-largest number of nodes belongs to the University of Queensland, at 22% of the total. The University of New South Wales follows them at 11.83%, Monash University at 4.8%, the University of Sydney at 4.53%, RMIT university at 2.67%, and the University of Melbourne at 2.26%. The rest of the universities make up the remainder of the total. In the clustering map, the University of Queensland and the University of New South Wales are depicted in the cluster's centre with interconnected nodes and generally are well represented.

The defence science institute is detected in clustering on the lower right end of the map. The authors are shown to interface predominately with the University of New South Wales at Canberra, and the University of Queensland. Other universities are generally disconnected from the co-authorships in hypersonic material from the institute. Within the cluster are a number of authors from the Boeing corporation, co-authoring papers regarding the Hi-Fire program. In the upper corner of the map is a dense, evenly spaced cluster, this happens to represent a paper concerning the luminosity of a stellar body. Although this is outside of hypersonic systems' general scope, the paper includes mention of hypersonic shockwaves and thus was included in the bibliographic data harvest. In a similar instance, another group, disconnected from the majority of papers in the central map, discusses hypersonic photons, and again is outside of the search's primary aims. Thus, using this particular clustering algorithm, papers in the centre of the map represent similar research fields. In this case, those fields include hypersonics for the means of propulsion, whereas separate groups on the fringes may investigate



a.) Publication frequency with respect to Australian states and territories



b.) Publications by volume from foreign countries

Figure 3: Geographic breakdown of hypersonic related publications

hypersonics outside of the intended purposes.

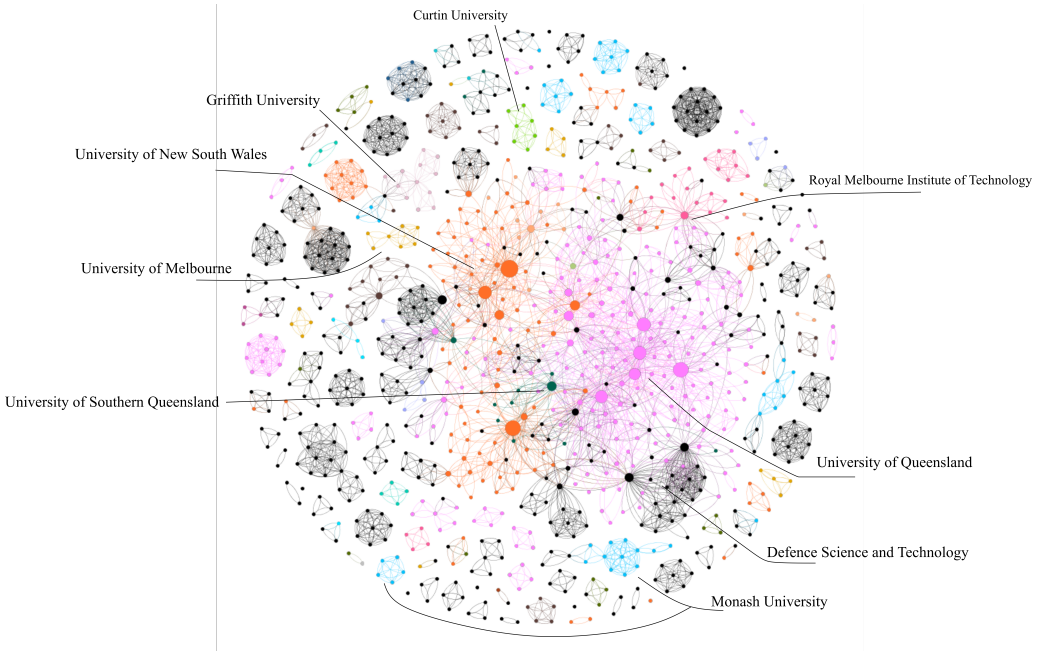


Figure 4: Co-authorship clustering

The full clustering map still includes over 800 authors, and analysing each author might be too time-consuming. But it is possible to identify key authors in the group, not through citations, or the number of publications, but due to their centrality within the co-authors. For example, Table 1 lists the key authors in the grouping with respect to university affiliation, ranked based on the normalised betweenness centrality. This depicts the top 17 central authors in the cluster, prominent members in a group and the first contact for engaging hypersonics research in this instance. Authors are designated A through to Q, with the highest betweenness centrality being sourced from an author at the University of New South Wales at the Australian Defence Force Academy, now known as the University of New South Wales at Canberra. The institution has significant influence in the field by being represented numerous times by authors A, B, E, and I. Followed in equal abundance by the University of Queensland as authors C, F, G, H, J, and Q. Notable additions include authors from the European Space Agency, Defence Science and Technology Group, University of Southern Queensland, RMIT

University, and the University of New South Wales at the Sydney based campus. Ultimately, these authors represent a select few who are critical or influential co-authors in the mapping network, which can be due to various reasons. From the standpoint of policy-driven action, these individuals represent the first point in canvassing technology initiatives for the keywords searched. The ease of constructing clustering maps through bibliographic data helps users identify groupings, dominant players in the area, and their associated connections.

Table 1: Key authors with respect to clustering centrality

	Institution	**Papers**	**Citations**	**Co-author**	**Rank**
A	University of New South Wales at Australian Defence Force Academy	4	71	31	1
B	University of New South Wales at Australian Defence Force Academy	2	1	36	0.86
C	University of Queensland	7	20	29	0.83
D	University of Queensland	5	35	23	0.71
E	University of New South Wales at Australian Defence Force Academy	7	9	29	0.70
F	University of Queensland	2	3	40	0.68
G	University of Queensland	7	49	33	0.63
H	University of Queensland	2	0	35	0.56
I	University of New South Wales at Australian Defence Force Academy	5	12	17	0.46
J	University of Queensland	2	3	6	0.44
K	University of Southern Queensland	3	2	22	0.42
L	UNSW Sydney	4	18	16	0.42
M	European Space Agency	3	35	13	0.40
N	Defence Science and Technology	3	11	25	0.39
O	Defence Science and Technology	6	119	29	0.36
P	RMIT University	5	36	21	0.29
Q	University of Queensland	2	1	13	0.27

In comparison to metrics about geography, the clustering shares specific observations. There is an abundance of researchers from New South Wales, the Australian Capitol Territory, and Queensland institutions. The clustering using the centrality metrics also shows that the United States and Europe are close contacts, as touched upon in papers by overseas authors.

However, the clustering approach with the centrality metrics does not necessarily highlight Singapore or Germany, in the top number of centrally placed contacts. Furthermore, the map itself shows that the Defence Science and Technology group tend to interact with the University of New South Wales and the University of Queensland. This could be an opportunity to interact with other similar research groups that are presently not directly connected such as those at the University of Southern Queensland, Griffith University, Monash University, and the Royal Melbourne Institute of Technology. Since hypersonics is an applied research area, domestic collaboration is less likely to occur unless framework is set in place to actively support collaboration due to competitive advance. Thus taking an approach that has been demonstrated in the United States with the clustering of universities under particular hypersonic research activities and centres of excellence.

3. Conclusion

The body of work examines the use of bibliographic clustering to aide in the canvasing of future technologies for policymakers. Using data-harvesting of the Scopus journal database and open-source software Gephi, it is readily possible for policy makers to examine research clusters in any field. This approach has been demonstrated using a few critical terms associated with hypersonic propulsion and has shown the dominant authors and institutions operating in this field. The clustering shows strong co-authorships between the University of New South Wales and the University of Queensland, indicating cooperative agreements or strong cross-collaborations. That said, various other researchers are also picked up through the betweenness centrality measures, including authors at the Defence Science Technology, European Space Agency, the University of Southern Queensland, and the Royal Melbourne Institute of Technology.

The canvasing demonstrated here also allows more detailed exploration of the data, which is visually easier to investigate. Thereby increasing productivity, and allowing those within the policy to examine the relationship between authors and minor entities. The work can potentially be applied to canvasing relevant staff and institutions for tendering, regarding relevant

technologies based on initiatives set out through federal and state government.

References

[1] H.-j. Lee, S. Lee, B. Yoon, Technology clustering based on evolutionary patterns: The case of information and communications technologies, Technological Forecasting and Social Change 78 (2011) 953–967.

[2] S. Valverde, R. V. Solé, M. A. Bedau, N. Packard, Topology and evolution of technology innovation networks, Physical Review E 76 (2007) 056118.

[3] R. Baptista, Geographical clusters and innovation diffusion, Technological Forecasting and Social Change 66 (2001) 31–46.

[4] S. Liyanage, Breeding innovation clusters through collaborative research networks, Technovation 15 (1995) 553–567.

[5] H. Dundar, D. R. Lewis, Determinants of research productivity in higher education, Research in higher education 39 (1998) 607–631.

[6] G. Abramo, C. A. D'Angelo, F. Di Costa, Research collaboration and productivity: is there correlation?, Higher education 57 (2009) 155–171.

[7] T. Velden, A.-u. Haque, C. Lagoze, A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation, Scientometrics 85 (2010) 219–242.

[8] R. Molontay, M. Nagy, Twenty years of network science: A bibliographic and co-authorship network analysis, arXiv preprint arXiv:2001.09006 (2020).

[9] C. Biscaro, C. Giupponi, Co-authorship and bibliographic coupling network effects on citations, PloS one 9 (2014) e99502.

[10] S. Murthy, E. Curran, Scramjet propulsion, American Institute of Aeronautics and Astronautics, 2001.

[11] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 3, 2009.

[12] F. P. Miller, A. F. Vandome, J. McBrewster, Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance (2009).