Finding Cross-Border Collaborative Centres in Biopharma Patent Networks: A Clustering Comparison Approach Based on Adjusted Mutual Information

Zhen Zhu[1] and Yuan Gao[2]

[1] University of Kent, Canterbury CT2 7NZ, UK
z.zhu@kent.ac.uk,
[2] University of East Anglia, Norwich NR4 7TJ, UK
y.gao4@uea.ac.uk

Abstract. The recent speedy development of COVID-19 mRNA vaccines has underlined the importance of cross-border patent collaboration. This paper uses the latest edition of the REGPAT database from the OECD and constructs the co-applicant patent networks for the fields of biotechnology and pharmaceuticals. We identify the cross-border collaborative regional centres in these patent networks at NUTS3 level using a clustering comparison approach based on adjusted mutual information (AMI). In particular, we measure and compare the AMI scores of the clustering before and after arbitrarily removing cross-border links of a focal node against the default clustering defined by national borders. The region with the largest difference in AMI scores is identified as the most cross-border collaborative centre, hence the name of our measure, AMI gain. We find that our measure both correlates with and has advantages over the traditional measure betweenness centrality and a simple measure of foreign share.

Keywords: patent networks, clustering comparison, adjusted mutual information, cross-border

1 Introduction

Globalisation and knowledge-based economy have stimulated the process of knowledge diffusion in the form of research and development (R&D) collaboration. Knowledge spillovers have been found to be geographically localised [1] and easier within firms than between [2]. R&D collaboration between organisations in different countries (across national borders or simply *cross-border* thereafter) could expose the participating parties to more heterogeneous resources, knowledge and skill sets. The data from the European Regional Innovation Survey from 1995 to 1997 has already shown that manufacturing firms with an external innovation network are more successful [3]. Conducting research on cross-border

knowledge diffusion is especially meaningful as R&D cooperation and dissemination of innovation have been identified as key indicators in the National Innovation System (NIS) studies [4, 5]. More recently, the development of COVID-19 mRNA vaccines on an unprecedented timescale has showcased the importance of cross-border patent collaborations [6]. In this paper, we focus on identifying regional centres in the cross-border collaborative networks as such centrality is associated with higher level of innovation intensity and quality. Our proposed identification method is based on the adjusted mutual information (AMI) gain by comparing each pair of elective partitions.

In quantitative innovation studies, patent information has been a widely used data source [7–12]. In the literature of R&D collaboration, researchers have been building linkages based on patent co-invention and co-application. In particular, the location information of patent inventors and applicants allows for accurate studies on cross-regional, co-inventionship and talent mobility. For example, Chessa et. al. constructed five networks using the OECD REGPAT database [13] to explore the R&D integration in the European Union. These include the patent co-inventor and publication co-author networks, the patent co-applicant network, the patent citation network and the patent inventor mobility network. Singh's analysis of patents filed to the U.S. Patent and Trademark Office (USPTO) uses patent citation data to measure the knowledge flow and builds interpersonal networks between inventors. In line with the previous literature like Kogut and Zander [2], this analysis shows intra-regional and intra-firm knowledge flows are stronger than those across regional or firm boundaries [14]. On the temporal dimension, a study based on patents originated from OECD countries and filed through the European Patent Office (EPO) found that the negative impact of geographical distance and institutional borders on R&D collaboration decreased from the end of 1980s till mid-1990s before it started to grow [15]. Further analysis looks into the how the quality of inter-regional knowledge networks (also based on the REGPAT patent database) impacts the regional research productivity [16]. REGPAT is also used in combination with the Eurostat database with a focus on the innovation-lagging-behind European regions to suggest that having wider inter-regional co-patenting networks with closer collaboration with knowledge-intensive regions could help the less innovative regions to close the gap [17].

As we have seen in the aforementioned literature, a rising number of literature have come to recognise the importance of knowledge spillovers. The earlier works look into various knowledge transmission channels (e.g., citation, collaboration, inventor mobility, etc), and the more recent studies began to leverage the power of network methods. But still, a relatively smaller body of literature have come up with a method to measure the regional R&D network centrality. So far the most common approaches derive from the conventional social network analysis (SNA), such as degree centrality or betweenness centrality [18, 19]. Berge et. al. argued that such studies could miss the conceptual problems at the aggregated regional level and lose the information regarding the structure of network relations [20]. They propose a new method based on the concept of inter-regional bridging

paths defined as the indirect connections between two regions via a third region as the bridge.

Our analysis conducts network construction based on the co-applicant linkages as they represent the collaboration between institutions. In terms of network centres identification, we take a different approach from the existing literature. Clustering comparison measures traditionally have been used for external validation as well as clustering solutions search [21]. In this paper, we propose another application of clustering comparison as a way of identifying central nodes in networks. In particular, we measure and compare the similarity scores of the clustering before and after arbitrarily removing cross-border links of a focal node against the default clustering defined by national borders. The widely used adjusted mutual information (AMI) is chosen here as the clustering comparison measure, hence the name of our measure, AMI gain. Using the examples of co-applicant patent networks in the fields of biotechnology and pharmaceuticals, we find that our measure, AMI gain, both correlates with and has advantages over the traditional measure of betweenness centrality and a simple measure of foreign share.

The rest of the paper is organised as follows: Section 2 introduces the database and our measure. Section 3 presents the results and statistically compares our measure with betweenness centrality and a simple measure of foreign share. Finally, Section 4 concludes the paper with further discussions.

2 Data and Methods

2.1 REGPAT Database

In this study, we use the latest edition of the OECD REGPAT database (released in January, 2021) which has been widely used in the relevant prior works. This database enables researchers to link patent data to regions based on the addresses of the patent applicants and inventors at NUTS3 level, covering more than 5,500 regions across OECD countries, EU 28 countries, Brazil, China, India, the Russian Federation and South Africa [13]. The patent data component in this database comes from the EPO Worldwide Statistical Patent Database (PATSTAT Global, Autumn 2020), which covers patent applications filed to the EPO and patent applications filed under the Patent Co-operation Treaty (PCT) at international phase, both from 1977 (priority date).

We focus the analysis on 30 countries in Europe, i.e., the EU28 countries except for Cyprus before the Brexit plus Iceland, Norway and Switzerland. As a result, we have 1389 NUTS3 level regions, i.e., the nodes in the networks. And a cross-border link occurs when it connects two regions belonging to two different countries. We construct two co-applicant patent networks for the two fields of biotechnology and pharmaceuticals according to the IPC concordance table published by the WIPO [22], where the nodes are the NUTS3 regions in these 30 countries and the links are weighted by the accumulated number of co-applicant collaboration instances between regions over time (i.e., from 1977

onward). Note that a patent may have one (i.e., contributing no links), two (i.e., contributing one link), or more (i.e., contributing more than one links) applicants. Also note that self-loops are considered and weighted. We further restrict our attention to the largest components of the two networks, with 765 nodes for biotechnology and 608 nodes for pharmaceuticals respectively.

2.2 Methods

We denote a network as $G = (V, E)$ where V is the set of nodes (or vertices) and E is the set of links (or edges). To describe our measure, we further denote $v_i \in V$ as node i in the network and $e_{v_i,v_j} \in E$ as the edge between node i and node j. The weight of e_{v_i,v_j} is denoted as w_{v_i,v_j} and $w_{v_i,v_j} = w_{v_j,v_i}$ for an undirected network. The set of node i's neighbouring (directly connected) nodes is denoted as $N(v_i)$. The largest component of the network is denoted as C_1. A partition i of the network is denoted as P_i. Finally, the partition after removing node i is denoted as P_{-v_i}. Regarding clustering comparison, we use adjusted mutual information (AMI), which calculates the similarity score between two partitions (or clusterings), say P_i and P_j, as follows:

$$AMI(P_i, P_j) = \frac{MI(P_i, P_j) - E\{MI(P_i, P_j)\}}{max(H(P_i), H(P_j)) - E\{MI(P_i, P_j)\}}$$

where $E\{\cdot\}$ calculates the expected value, $H(\cdot)$ calculates the entropy and $MI(\cdot)$ calculates the (unadjusted) mutual information [21]. The value of AMI ranges from 0 to 1 and 0 implies the most dissimilar whereas 1 implies the most similar between partitions.

Algorithm 1 shows the pseudocode of calculating the AMI gain for each node. Note that for each node we conduct a counterfactual exercise by arbitrarily removing its cross-border links. The rationale behind our measure is that such a counterfactual exercise will produce a partition more similar to the default partition defined by national borders, for which we denote as P_d. Therefore, the difference between the AMI scores when compared with the default partition will more than often be positive after the node removal and we call the difference as the AMI gain. As a result, the cross-border collaborative centres are identified with the largest AMI gains. Note that we use the Louvain method [23] at the default resolution level 1.0 for community detection, which also takes into account link weights.

Algorithm 1 Calculating AMI gain

$P_0 \leftarrow Louvain(C_1)$ ▷ Get P_0 by applying Louvain to the largest component C_1
$AMI_0 \leftarrow AMI(P_0, P_d)$ ▷ AMI between P_0 and the default partition P_d
for $v_i \leftarrow v_1, v_n$ **do** ▷ Loop through the nodes of C_1
  for $N(v_i)_j \leftarrow N(v_i)_1, N(v_i)_m$ **do** ▷ Loop through the neighbours of v_i
    if $N(v_i)_j$ is cross-border **then**
      remove $e_{v_i,N(v_i)_j}$ ▷ Drop cross-border neighbours of v_i
    end if
  end for
  $P_{-v_i} \leftarrow Louvain(C_{-v_i})$
  $AMI_{-v_i} \leftarrow AMI(P_{-v_i}, P_d)$
  $\Delta AMI_{v_i} = AMI_{-v_i} - AMI_0$ ▷ AMI gain for node v_i
end for

For comparison, we also consider the traditional measure of betweenness centrality (which also takes into account link weights) and a simple measure of foreign share. Algorithm 2 shows the pseudocode of calculating the foreign share for each node.

Algorithm 2 Calculating foreign share

for $v_i \leftarrow v_1, v_n$ **do** ▷ Loop through the nodes of C_1
  $sum_{v_i} \leftarrow 0$
  $sum^f_{v_i} \leftarrow 0$
  for $N(v_i)_j \leftarrow N(v_i)_1, N(v_i)_m$ **do** ▷ Loop through the neighbours of v_i
    if $N(v_i)_j$ is cross-border **then**
      $sum^f_{v_i} \leftarrow sum^f_{v_i} + w_{v_i,N(v_i)_j}$ ▷ Add up foreign neighbour edge weights
      $sum_{v_i} \leftarrow sum_{v_i} + w_{v_i,N(v_i)_j}$ ▷ Add up total neighbour edge weights
    else
      $sum_{v_i} \leftarrow sum_{v_i} + w_{v_i,N(v_i)_j}$ ▷ Add up total neighbour edge weights
    end if
  end for
  $FS_{v_i} = \frac{sum^f_{v_i}}{sum_{v_i}}$ ▷ Compute foreign share for node v_i
end for

3 Results

As a concrete example, Figure 1 and Figure 2 show the community detection results before and after we arbitrarily remove the cross-border links of the region BE234 (Arr. Gent) from the biotechnology patent network. Each color represents a different community detected. Note that mostly the communities are still characterised by national borders, even though cross-border links sometimes break certain regions from their default national communities (e.g., different colors within France and Germany). Also note that a significant difference between the two figures is that most regions in Netherlands share the community with the

UK (Figure 1) before the counterfactual removal but are separated from the UK (Figure 2) after the removal, which helps BE234 (Arr. Gent) attain a high AMI gain score.

Table 1 shows the top 10 regions identified by our measure, AMI gain, as well as by betweenness centrality in the field of biotechnology. Although not shown in the table, foreign share has identified 32 regions all with 100% cross-border connections in the field of biotechnology, including, for example, Kelheim in Germany and Malta. Similarly, Table 2 shows the top 10 regions identified by our measure, AMI gain, versus by betweenness centrality, in the field of pharmaceuticals. Again not shown in the table, foreign share has identified 37 regions in tie in the field of pharmaceuticals, including, for example, Plymouth in the UK and Malta.

There is some overlapping between the results by AMI gain and by betweenness centrality. For example, Vienna and Copenhagen in Table 1, and Stockholm, Milan and Paris in Table 2. On the other hand, the local measure of foreign share cannot differentiate the regions very well on the top as many regions are in tie. Moreover, foreign share does not take into account structural and global properties of the network. A small region, such as Malta, simply with all its links cross-border would top the table by foreign share.

Fig. 1. Biotechnology patent network community detection result

Fig. 2. Biotechnology patent network community detection result (cross-border links removed for BE234)

Table 1. Top 10 regions in biotechnology

	AMI gain		**Betweenness centrality**	
#	NUTS3 code	Region name	NUTS3 code	Region name
1	BE234	Arr. Gent	FR101	Paris
2	DE126	Mannheim Stadtkreis	CH031	Basel-Stadt
3	UKD31	Greater Manchester South	DE212	München Kreisfreie Stadt
4	DEE02	Halle (Saale) Kreisfreie Stadt	UKI11	Inner London - West
5	AT130	Vienna	ITI43	Rome
6	DK011	City of Copenhagen	ES300	Madrid
7	SE110	Stockholm County	SE110	Stockholm County
8	UKF22	Leicestershire CC and Rutland	UKJ14	Oxfordshire
9	DE125	Heidelberg Stadtkreis	DE300	Berlin
10	AT323	Salzburg und Umgebung	AT130	Vienna

Table 2. Top 10 regions in pharmaceuticals

	AMI gain		**Betweenness centrality**	
#	NUTS3 code	Region name	NUTS3 code	Region name
1	UKJ33	Hampshire CC	FR101	Paris
2	SE110	Stockholm County	UKH12	Cambridgeshire CC
3	DK011	City of Copenhagen	CH031	Basel-Stadt
4	DEA22	Bonn Kreisfreie Stadt	SE110	Stockholm County
5	DEA2B	Rheinisch-Bergischer Kreis	CH011	Vaud
6	ITC4C	Milan	ES300	Madrid
7	FR101	Paris	DE300	Berlin
8	DE926	Holzminden	ITC4C	Milan
9	ITC33	Genoa	AT130	Vienna
10	ITG2C	Carbonia-Iglesias	DE125	Heidelberg Stadtkreis

More systematically, Figure 3 shows the scatter plots between our measure, AMI gain, and betweenness centrality or foreign share for biotechnology and pharmaceuticals respectively. The Pearson correlation coefficient (denoted by r) as well as the Spearman correlation coefficient (denoted by ρ) between AMI gain and either of the two alternative measures are positive. Note that the Spearman correlations are stronger than the Pearson ones as the former only considers the ranking of the values. Therefore, our measure, AMI gain, captures certain similar information as either betweenness centrality or foreign share does but also differs from either of them in a nontrivial way.

Furthermore, Figure 4 shows the empirical cumulative distribution functions (ECDFs) of our measure, AMI gain, betweenness centrality and foreign share for biotechnology and pharmaceuticals respectively. For both fields, betweenness centrality results are dominated by a few regions (as its ECDF curve is bent towards the top left corner). AMI gain and foreign share have, relatively speaking, more uniform distributions of values (i.e., closer to the 45 degree line). As a result, AMI gain helps identify and differentiate the central cross-border collaborative regions on the top (better than foreign share) and for a large percentile range (better than betweenness centrality).

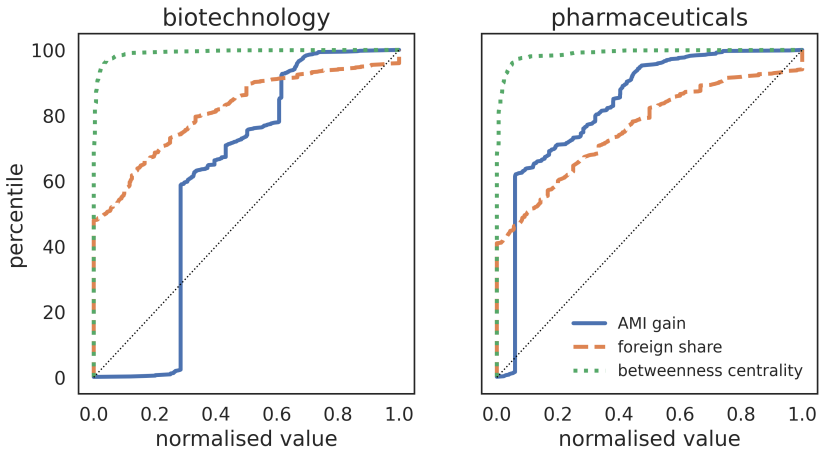

Fig. 4. Empirical cumulative distribution functions of the three measures

4 Conclusions

R&D collaborations beyond national borders are critical for knowledge spillovers at large scale, which is well demonstrated by the recent development of COVID-19 mRNA vaccines at an unprecedented timescale. This paper uses the latest edition of the REGPAT database from the OECD and constructs the co-applicant

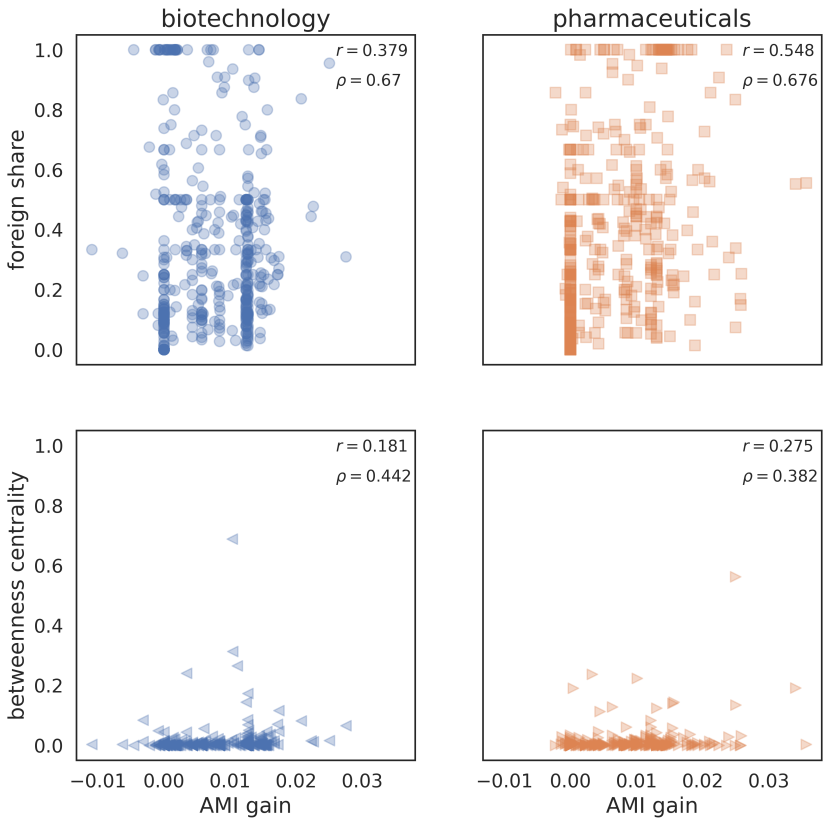

Fig. 3. Correlations between the three measures

patent networks in Europe at NUTS3 level for the fields of biotechnology and pharmaceuticals.

We contribute to the literature of finding cross-border collaborative centres in patent networks by proposing a clustering comparison approach based on adjusted mutual information. The rationale behind our approach is that a counterfactual exercise of removing cross-border links from a focal node will produce a partition more similar to the default partition defined by national borders. Therefore, the difference between the AMI scores when compared with the default partition will more than often be positive after the node removal. The results based on our measure, AMI gain, are positively correlated with those by betweenness centrality or by a simple measure of foreign share. Nevertheless, when compared with betweenness centrality, AMI gain better differentiates cross-border centres from local ones and offers a more uniform distribution of values. On the other hand, when compared with foreign share, AMI gain is more of a global and structural measure and better differentiates the nodes on the

top. Our future research will further explore the robustness of our measure with more variations of the parameters and across contexts.

References

1. Adam B Jaffe, Manuel Trajtenberg, and Rebecca Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598, 1993.
2. Bruce Kogut and Udo Zander. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization science*, 3(3):383–397, 1992.
3. Knut Koschatzky and Rolf Sternberg. R&d cooperation in innovation systems—some lessons from the european regional innovation survey (eris). *European planning studies*, 8(4):487–501, 2000.
4. Organisation for Economic Co-operation and Development (OECD). *Managing national innovation systems*. OECD Publishing, 1999.
5. Pao-Long Chang and Hsin-Yu Shih. The innovation systems of taiwan and china: a comparative analysis. *Technovation*, 24(7):529–539, 2004.
6. Mario Gaviria and Burcu Kilic. A network analysis of covid-19 mrna vaccine patents. *Nature Biotechnology*, 39(5):546–548, 2021.
7. Zvi Griliches, Ariel Pakes, and Bronwyn H Hall. The value of patents as indicators of inventive activity, 1986.
8. Lee Fleming. Recombinant uncertainty in technological search. *Management science*, 47(1):117–132, 2001.
9. Adam B Jaffe and Manuel Trajtenberg. *Patents, citations, and innovations: A window on the knowledge economy*. MIT press, 2002.
10. Bronwyn H Hall, Adam Jaffe, and Manuel Trajtenberg. Market value and patent citations. *RAND Journal of economics*, pages 16–38, 2005.
11. Yuan Gao, Zhen Zhu, and Massimo Riccaboni. Consistency and trends of technological innovations: A network approach to the international patent classification data. In *International Conference on Complex Networks and their Applications*, pages 744–756. Springer, 2017.
12. Yuan Gao, Zhen Zhu, Raja Kali, and Massimo Riccaboni. Community evolution in patent networks: technological change and network dynamics. *Applied network science*, 3(1):1–23, 2018.
13. Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec. The oecd regpat database: a presentation. *OECD Science, Technology and Industry Working Papers*, 2008(2):0_1, 2008.
14. Jasjit Singh. Collaborative networks as determinants of knowledge diffusion patterns. *Management science*, 51(5):756–770, 2005.
15. Andrea Morescalchi, Fabio Pammolli, Orion Penner, Alexander M Petersen, and Massimo Riccaboni. The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy*, 44(3):651–668, 2015.
16. Tamás Sebestyén and Attila Varga. Research productivity and the quality of interregional knowledge networks. *The Annals of Regional Science*, 51(1):155–189, 2013.
17. Ivan De Noni, Luigi Orsi, and Fiorenza Belussi. The role of collaborative networks in supporting the innovation performances of lagging-behind european regions. *Research Policy*, 47(1):1–13, 2018.

18. Iris Wanzenboeck, Thomas Scherngell, and Thomas Brenner. Embeddedness of regions in european knowledge networks: a comparative analysis of inter-regional r&d collaborations, co-patents and co-publications. *The Annals of Regional Science*, 53(2):337–368, 2014.
19. Iris Wanzenböck, Thomas Scherngell, and Rafael Lata. Embeddedness of european regions in european union-funded research and development (r&d) networks: A spatial econometric perspective. *Regional Studies*, 49(10):1685–1705, 2015.
20. Laurent R Bergé, Iris Wanzenböck, and Thomas Scherngell. Centrality of regions in r&d networks: A new measurement approach using the concept of bridging paths. *Regional Studies*, 51(8):1165–1178, 2017.
21. Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.
22. WIPO. Ipc concordance table. https://www.wipo.int/ipstats, 2019. Last accessed 2021-08-06.
23. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.