Identifying research topics and collaboration networks in Finland: topic modelling of scientific publications in 2008–2019

Katja Mankinen and Yrjö Leino

ACADEMY OF FINLAND

ISBN 978-951-715-925-8

Tiivistelmä

Tämän työn tavoitteena oli datalähtöisesti kartoittaa Suomessa tutkittavia aiheita ja ilmiöitä Web of Science -tietokannassa olevien vuosina 2008—2019 julkaistujen tieteellisten julkaisujen pohjalta. Kokeilevan ja alustavan analyysin tarkoituksena oli selvittää, voiko tekstinlouhintaa ja bibliometrisiä menetelmiä käyttää ilmiöpohjaisten tutkimusaiheiden ja niihin liittyvän tutkimusyhteistyön tunnistamiseen. Työ on tehty yhteistyössä Suomen Akatemian kanssa Tieteen Tila -analyysien yhteydessä.

Web of Science Core Collection -tietokannasta haettiin yhteensä 106736 englanninkielistä vuosina 2008—2019 julkaistua julkaisua (artikkelia, katsausartikkelia, kirjettä lehden toimituskunnalle, kirjaa), joissa on vähintään yksi tekijä suomalaisesta affiliaatiosta. Aiheet tunnistettiin julkaisujen otsikkoon, tiivistelmiin ja avainsanoihin perustuen koneoppimiseen pohjautuvalla aihemallinnusmenetelmällä. Aiheet siis määräytyivät datan ja algoritmien eivätkä ennalta määriteltyjen luokkien tai tieteenalojen perusteella. Tavallista todennäköisyyspohjaista Latent Dirichlet Allocation -aihemallinnusmenetelmää parannettiin sanavektoreilla semanttisen ymmärryksen lisäämiseksi.

Kaikkiaan aihemallinnusohjelmalla generoitiin 1026 aihetta. Aiheisiin liittyvää organisaatioiden yhteistyötä tutkittiin rakentamalla yhteisjulkaisuverkostoja, jotka kuvaavat eri organisaatioiden välistä yhteistyötä kussakin aiheessa. Tieteellistä vaikuttavuutta arvioitiin tarkastelemalla eniten viitattujen julkaisujen suhteellista osuutta aiheittain. Lisäksi tarkasteltiin julkaisumäärien kehitystä eri aiheissa vuosina 2008—2019 mahdollisten tutkimustrendien löytämiseksi.

Viittausmäärien perusteella vaikuttavaa tutkimusta monessa eri aiheessa

Julkaisujen lukumäärä aiheittain vaihtelee 18 ja 862 välillä. Julkaisumäärän keskiarvo on 104.0 ja mediaani 81.5. Eniten julkaisuja sisältävät aiheet edustavat laajoja tieteenalamaisia kokonaisuuksia koulutuksesta liiketoimintaan ja liikkeenjohtoon, langattomiin verkkoihin, uusiutuvaan energiaan ja suolistomikrobistoon, kun taas pienimmät aiheet ovat yleensä hyvin spesifisiä, kuten erillisiä eläin- tai kasvilajeja tai sairauksia. Samaa yleistä teemaa voidaan käsitellä monissa aiheissa eri näkökulmista. Kaikki aiheeseen luokitellut julkaisut eivät kuitenkaan välttämättä liity tiukasti kyseiseen aiheeseen, eikä aihe sisällä jokaista kyseistä aihetta koskevaa julkaisua. Tästä syystä aiheen julkaisujen määrä ei täysin vastaa ilmiön todellista kokoa. Lisäksi aihemallinnusalgoritmi tuotti muutamia pieniä, heikkolaatuisia aiheita, joissa julkaisujen välillä ei ole selvää yhteyttä.

Viittausmääriltään vaikuttavien tutkimusaiheiden löytämiseksi laskettiin top 10 -

indeksi aiheittain. Aiheiden joukosta löydettiin 65 vaikuttavaa aihetta, joiden top 10 -indeksi on yli 2.0, ja 10 hyvin vaikuttavaa aihetta, joiden top 10-indeksi on yli 3.5. Suurin osa näistä korkean indeksin aiheista liittyy tietojenkäsittelytieteen, sähkötekniikan, ympäristötieteiden, materiaalitieteen, tietoliikenteen ja johtamisen eri osa-alueisiin.

Enemmän kansainvälistä kuin kansallista yhteistyötä korkean top 10 -indeksin aiheissa

Myös kansallista ja kansainvälistä julkaisuyhteistyötä tutkittiin aiheittain. Yleisesti viittausmääriltään vaikuttavien tutkimusaiheiden joukossa oli enemmän kansainvälisiä kuin kansallisia yhteisjulkaisuja.

Joissakin aiheissa kansainvälinen yhteisjulkaiseminen korostui ja kansallinen yhteisjulkaiseminen oli lähes poikkeus. Tällaisia olivat monet tietoliikenteeseen ja sähkötekniikkaan liittyvät aiheet. Useissa muissa aiheissa, kuten laserkeilaukseen liittyvässä aiheessa, kansallinen yhteistyö oli laajaa ja enimmäkseen viittausmääriltään vaikuttavaa. Aiheen vaikuttavuus voi olla myös vahvasti polarisoitunut: kaikista aiheessa aktiivisista organisaatioista ehkä vain muutama tuottaa erityisen vaikuttavaa tutkimusta. Korkea kansallinen top 10 -indeksi ei siis suoraan tarkoita, että kaikkialla Suomessa tehtäisiin vaikuttavaa tutkimusta aiheesta. On kuitenkin huomattava, että jonkin organisaation matala top 10 -indeksi voi johtua monista syistä, esimerkiksi väärään aiheeseen luokitelluista julkaisuista.

Merkittäviäkin tutkimusaiheita voi jäädä löytymättä

Vaikka tämän kokeellisen projektin tarkoituksena oli tunnistaa tutkimusaiheita, merkittäviäkin aiheita voi jäädä tunnistamatta aineiston ja valittujen menetelmien heikkouksien vuoksi. Esitetty työ ei pysty määrittämään tutkimusaiheita ja niihin liittyviä organisaatioita yksikäsitteisesti. Vaikka käytetyt menetelmät soveltuvat tutkimusaiheiden kartoittamiseen, ne eivät sovi tarkkaan tutkimusaiheiden ja -ilmiöiden koon määrittämiseen tai luokitteluun.

Valittu aineisto on vain osajoukko kaikesta vuosina 2008—2019 julkaistusta suomalaisesta tutkimuksesta. Aineiston kattavuus on heikko erityisesti humanistisissa tieteissä ja yhteiskuntatieteissä. Datalähtöinen aihemallinnusmenetelmä ei ota huomioon julkaisun tieteenalaa, viittauksia tai ontologiapohjaista luokittelua aiheen tunnistamisen parantamiseksi. Lisäksi menetelmässä asetetaan lukuisia parametreja, joiden muuttaminen voi johtaa hyvin erilaisiin aiheisiin ja julkaisujoukkoihin.

Tutkimusaiheiden perusteellisemmassa tarkastelussa kannattanee käyttää laajempaa aineistoa (mukaan lukien suomen- ja ruotsinkieliset julkaisut ja muut

julkaisutyypit) ja ottaa julkaisujen viittaustiedot huomioon. Tutkimusaiheiden taloudellista vaikuttavuutta voitaisiin arvioida esimerkiksi patenttitietojen perusteella. Osaamiskeskittymien tunnistamiseksi ja analysoimiseksi tarvitaan yhteisjulkaisuverkostojen lisäksi myös muuta laadullista aineistoa.

Nämä rajoitukset huomioiden tämä koneoppimismenetelmiä ja bibliometrisiä analyysejä yhdistävä raportti tarjoaa datalähtöisiä näkemyksiä suomalaisesta tutkimuksesta.

Sammanfattning

Syftet med denna rapport var att på ett databaserat sätt kartlägga teman och fenomen som undersöks i Finland utifrån uppgifter i databasen Web of Science över vetenskapliga publikationer som publicerats åren 2008–2019. Syftet med den experimentella och preliminära analysen var att utreda om textutvinning och bibliometriska metoder kan användas för att identifiera fenomenbaserade forskningsteman och forskningssamarbete i anslutning till dem. Arbetet utfördes i samarbete med Finlands Akademi och i samband med analyser om vetenskapens tillstånd i Finland.

Ur databasen Web of Science Core Collection söktes sammanlagt 106 736 publikationer på engelska (artiklar, översiktsartiklar, brev till redaktioner, böcker) som publicerats åren 2008–2019 och som innehöll minst en skribent med en finländsk bakgrundsorganisation. Temana identifierades utifrån publikationernas rubrik, sammanfattningar och nyckelord med hjälp av en ämnesmodellering som baserar sig på maskininlärning. Ämnena fastställdes alltså på grundval av data och algoritmer och inte utifrån på förhand fastställda kategorier eller vetenskapsgrenar. Den vanliga sannolikhetsbaserade metoden för ämnesmodellering, Latent Dirichlet Allocation, förbättrades med ordvektorer för att öka den semantiska förståelsen.

Ämnesmodelleringsprogrammet genererade sammanlagt 1 026 teman. Organisationernas samarbete kring temana undersöktes genom att man byggde upp sampubliceringsnätverk som beskriver samarbetet mellan olika organisationer i varje tema. Det vetenskapliga genomslaget bedömdes genom en ämnesvis granskning av den relativa andelen av de mest citerade publikationerna. Dessutom granskades utvecklingen av publikationsvolymerna i olika teman åren 2008–2019 för att identifiera eventuella forskningstrender.

Antalet citeringar visar på effektfull forskning inom många teman

Antalet publikationer varierar mellan 18 och 862 inom olika ämnen. Medeltalet för publikationsvolymen är 104.0 och medianvärdet 81.5. De teman som innehåller flest publikationer representerar omfattande disciplinliknande helheter från utbildning till affärsverksamhet och företagsledning, trådlösa nät, förnybar energi och tarmmikrober, medan de minsta temana i allmänhet är mycket specifika, såsom specifika djur- eller växtarter eller sjukdomar. Samma allmänna tema kan behandlas ur olika synvinklar i många ämnen. Alla publikationer som har klassificerats till ett visst ämne har dock inte nödvändigtvis ett strikt samband med ämnet i fråga, och ämnet omfattar inte alla publikationer som gäller ämnet i fråga. Därför är antalet publikationer i ämnet inte helt i linje med fenomenets verkliga storlek. Dessutom producerade modelleringsalgoritmen några små ämnen av dålig kvalitet där det inte finns någon klar koppling mellan publikationerna.

Topp 10-indexet fastställdes enligt ämnesområde för att man skulle kunna identifiera ämnen med ett stort antal citeringar. Bland ämnena identifierades 65 verkningsfulla ämnen med ett topp 10-index över 2.0 och tio mycket verkningsfulla ämnen med ett topp 10-index över 3.5. De flesta av dessa ämnen med ett högt indextal rör olika delområden inom datavetenskap, elektroteknik, miljövetenskap, materialvetenskap, telekommunikation och ledning.

Mer internationellt än nationellt samarbete kring topp 10-indexets teman

Även det nationella och internationella publikationssamarbetet undersöktes ämnesvis. I allmänhet fanns det fler internationella än nationella sampublikationer bland de forskningsteman som hade stort antal citeringar.

I vissa frågor accentuerades internationell sampublicering och den nationella sampubliceringen var nästan ett undantag. Det handlade bl.a. om många teman som gäller datakommunikation och elteknik. I flera andra teman, t.ex. laserskanning, var det nationella samarbetet omfattande och till största delen imponerande i fråga om antal citeringar. Temats genomslagskraft kan också vara starkt polariserad: av alla organisationer som är aktiva inom ett tema producerar kanske endast ett fåtal särskilt genomslagskraftig forskning. Ett högt nationellt topp 10-index innebär alltså inte direkt att det i hela Finland bedrivs genomslagskraftig forskning i ämnet. Det bör dock noteras att en organisations låga topp 10-index kan bero på många orsaker, t.ex. publikationer som klassificerats till fel tema.

Även betydande forskningsteman kan bli oupptäckta

Även om syftet med detta experimentella projekt var att identifiera forskningsteman, kan även viktiga teman bli oidentifierade på grund av brister i materialet och de valda metoderna. Det går inte att inte entydigt definiera forskningstemana och de därtill hörande organisationerna. Även om de metoder som använts lämpar sig för att kartlägga forskningsämnen lämpar de sig inte för att noggrant fastställa eller klassificera ämnenas och fenomenens storlek.

Det valda materialet utgör endast en del av all finländsk forskning som publicerats åren 2008–2019. Materialets täckning är låg i synnerhet inom humaniora och samhällsvetenskaper. Databaserad ämnesmodellering beaktar inte publikationens vetenskapsgren, citeringar eller ontologibaserade klassificering för att förbättra identifieringen av ämnet. Dessutom fastställer metoden ett stort antal parametrar som vid ändringar kan leda till mycket olika teman och publikationsgrupper.

Vid en grundligare granskning av forskningstemana lönar det sig sannolikt att använda mer omfattande material (inklusive finsk- och svenskspråkiga publikationer och andra publikationstyper) och beakta publikationernas citeringsuppgifter. Forskningstemanas ekonomiska effekter kunde bedömas t.ex. på basis av patentuppgifter. För att identifiera och analysera kompetenskluster behövs utöver sampublikationsnätverk även annat kvalitativt material.

Med beaktande av dessa begränsningar erbjuder denna rapport, som kombinerar maskininlärningsmetoder och bibliometriska analyser, databaserade synpunkter på finländsk forskning.

Executive summary

The aim of this project was to identify research topics in Finland based on Web of Science publication data from 2008 to 2019. The analysis was exploratory in nature and aimed at determining whether automated text mining and bibliometric methods can be used to find phenomenon-based research topics and associated research collaborations. The project was undertaken jointly with the Academy of Finland in the context of its State of Scientific Research analyses.

We searched the Web of Science Core Collection and retrieved 106,736 English-language publications (articles, reviews, letters, books) with at least one author with a Finnish affiliation and published in 2008–2019. To identify research topics and trends in Finland in a fully data-driven and unsupervised way, natural language

processing and topic modelling methods were applied to the abstracts, titles and keywords of research papers. This way, the topics are determined by the data and algorithms rather than by predetermined categories or scientific disciplines. A traditional probabilistic topic modelling method, Latent Dirichlet Allocation, was enhanced by word embeddings to capture the semantic knowledge at the publication level.

In total, 1,026 research topics were generated. Collaborations of organisations were sought by building co-authorship networks that describe cooperation between different organisations in each topic. Citation impact was evaluated by considering the share of highly cited publications within each topic. In addition, the development of publication volumes by topics in 2008–2019 was studied in order to detect possible research trends.

Research with high citation impact on many different topics

The number of publications in each topic identified varies between 18 and 862, with the mean being 104.0 and the median 81.5. Topics with the most publications represent broad scientific disciplines from education and business management to wireless networks, renewable energy and gut microbiota, whereas the smallest topics tend to be very specific such as distinct animal or plant species or diseases. The same general theme can be addressed in many topics from different perspectives. However, not all publications in the topic are necessarily strictly related to the topic, and not every publication concerning the topic in question has been clustered into that topic. Therefore, the number of publications on the topic does not fully reflect the true size of the phenomenon. In addition, the topic modelling algorithm generated a few small low-quality categories in which no clear connections can be found between more than a couple of publications.

In order to find high-impact research, top 10 indices were calculated for publications by topic. In total 65 topics with high citation impact (top 10 index over 2.0) and 10 topics with very high citation impact (top 10 index over 3.5) were found. Most of the topics with the highest top 10 indices are related to various aspects of computer science, electrical engineering, environmental sciences, materials science, telecommunications and management.

More international than national collaboration in topics with high citation impact

For each topic, national and international co-authored publications were studied and their citation impact evaluated. In general, there was more international than national collaboration in the topics with high citation impact.

Some topics showed an emphasis towards international collaboration while national collaboration was almost an exception. This was true of many topics related to telecommunications and electrical engineering, for example. In many other topics, such as laser scanning, there was extensive national collaboration and citation impacts were mainly high. Citation impacts can also be strongly polarised: among all the organisations active in a given topic, it is possible that only a few are producing high-impact research. Therefore, a high national citation impact does not directly imply that research in Finland is high-impact across the board. However, it is noteworthy that if an organisation is found to produce lower-impact research, that may be due to various reasons, such as misclassified publications.

Even significant research topics may go undiscovered

Although the purpose of this exploratory project was to identify research topics with high citation impact, it is possible that even significant topics remained unrecognized due to limitations of the dataset and weaknesses of the method applied. The project is unable to provide unambiguous definitions of research topics and related organisations. While the methods employed are suitable for mapping research topics at scale, they cannot be used for more accurate classifications or for specifying the exact size of research areas.

The selected data sample is only a subset of all Finnish research published in 2008–2019. Its coverage is especially low in the humanities and social sciences. The data-driven topic modelling method is unsupervised and does not take account of the field of the publication, citations or ontology-based classification in order to improve topic identification. In addition, the method involves setting numerous parameters, and any changes to those parameters can lead to very different topics and sets of publications.

For a more in-depth study of the topics and related research ecosystems, it would probably make sense to use a wider range of data (including other publication types and publications written in Finnish and Swedish) and also to take account of citation information. Economic impact could be assessed by means of patent data, for example. While the presented analysis shows a data-driven way to find collaboration networks related to the research topics identified, other qualitative material is needed to recognize and analyse real research ecosystems.

Given these limitations, this report provides data-driven insights into Finnish research by combining machine learning methods and bibliometric analyses.

1. Introduction

One of the key objectives of Finland's National Roadmap for Research, Development and Innovation (RDI) (2020) is to strengthen and support research ecosystems. Likewise, the Academy of Finland's report on the State of Scientific Research (2018) emphasizes that research ecosystems together with multidisciplinary and phenomenon-based research are major drivers of change in Finnish science. But how to identify such research and research ecosystems? Bibliometric analyses are traditionally conducted using predefined discipline categories. With the continuing growth of multidisciplinary and phenomenon-based research, that research will not fall neatly into traditional classification schemes. On the one hand, publications from a single phenomenon-based topic may appear in journals representing several different subject fields; on the other hand, a topic may be so subtle or specific that it remains undetected in usual bibliometric analyses.

This report describes a preliminary, experimental study that aims to identify research topics and related research collaboration ("research ecosystems") in Finland in a purely data-driven manner based on scholarly publications and co-authorship networks. Research topics are found by using unsupervised natural language processing methods, in particular topic modelling, on the titles, abstracts and keywords of scientific publications retrieved from the Web of Science (WoS) Core Collection. In this work, research ecosystems refer to groups of organisations that co-author highly cited publications on the topics. These ecosystems are identified from co-authorship networks. Therefore, an ecosystem is not necessarily any officially formed or funded group of researchers. Ecosystems are recognized only at the organisation level (such as universities) and not at the department or research unit level. Finally, the citation impact of topics is evaluated using the top 10 citation index.

1.1. Related work

The classification of publications into research areas is essential for bibliometric analyses and performance evaluation (Waltman & van Eck, 2012). In journal-defined classification systems, such as the Web of Science subject categories, publications are assigned to a research field based on the journal in which it is published. However, this approach creates problems with multidisciplinary journals such as Science and Nature (Waltman & van Eck, 2012). Another way is to aggregate publications directly at the publication level. One example is clustering publications into research fields based on their citation relationships, such as bibliographic couplings and co-citations (Traag, et al., 2019; Waltman & van Eck, 2012).

Publications can also be clustered into coherent research areas using their textual content, such as titles, abstracts and even full texts.

In machine learning and text mining, topic modelling is a technique that aims to discover underlying topics in a large volume of data. Topic modelling methods are usually unsupervised: instead of mapping publications to predefined categories,[1] the aim is to infer latent topics occurring in publications and to find characteristic word distributions for each topic. A topic is essentially a set of words that are likely to appear together in the same context: for example, a publication about politics may include words such as “election”, “parliament”, “vote”, and “president”, but no prior knowledge of such words or topics involved is needed. A major advantage of topic models is that publications can indeed be grouped together based on their topics instead of strictly predefined fields of science. For example, a broad topic such as “climate change” can be studied in many fields from ecology and biology to social sciences.

Several topic modelling methods have been proposed to recognize main topics in a collection of documents. These include non-probabilistic linear-algebraic methods such as Latent Semantic Indexing (Deerwester, et al., 1990), which uses a Singular Value Decomposition to decompose high-dimensional term-document matrices into lower-dimensional representations of topics. Glenisson et al. (2005) applied the Latent Semantic Indexing method to scholarly publications. Non-negative Matrix Factorisation (Paatero & Tapper, 1994) performs a similar decomposition task by setting non-negative constraints on the factorisation of matrices.

In addition to linear-algebraic methods, topic modelling has used several generative probabilistic approaches, including Probabilistic Latent Semantic Indexing (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, et al., 2003). The idea behind probabilistic models is to find latent topics that are “hidden”, as the only observed variables are the words in publications. In particular, the Latent Dirichlet Allocation (LDA) model has often been used to identify topics in large publication datasets. For example, Griffiths and Steyvers (2004) applied LDA to understand fields of science using the Proceedings of the National Academy of Sciences. LDA represents each publication as a mixture of topics and each topic as a distribution of words. Publications are processed as a “bag of words”: it is assumed that the words in a publication occur independently and that their order does not matter. The probabilities of words and publications belonging to a certain topic are estimated using Dirichlet distributions for each word and each publication.

[1] Categories could include hierarchically-organised, domain-specific vocabularies such as Physics Subject Headings (PhySH) and Medical Subject Headings (MeSH), or other well-established classification schemes of research fields, such as the Organisation for Economic Cooperation and Development (OECD) and the Web of Science subject categories.

There are also several extensions to Latent Dirichlet Allocation, such as the Author-Topic model (Rosen-Zvi, et al., 2004) in which each author is associated with a multinomial distribution over topics. Author-Conference-Topic (Tang, et al., 2008) simultaneously relates topics to publications, authors and conferences. Apart from static models, there are dynamic topic models which capture the evolution of topics over time (Blei & Lafferty, 2006). In these parametric models, the number of topics has to be pre-specified. In non-parametric models, such as the Bayesian Hierarchical Dirichlet Process model (Teh, et al., 2006), the number of topics is automatically inferred. Gerlach et al. (2018) proposed a network approach to topic models, where text corpora are presented as bipartite networks of documents and words, and topics are realized by community detection methods.

One of the main limitations of traditional topic modelling methods is that they are built on the “bag of words” approach, which does not take account of the semantic relation between words. To overcome this limitation, models can be extended to incorporate bigrams, i.e., pairs of words (Wallach, 2006). Nguyen et al. (2015) improved LDA in the LF-LDA model by using pre-trained word vector representations word2vec (Mikolov, et al., 2013) and GloVe (Pennington, et al., 2014). In this report, we use a similar methodology; the details of implementation are discussed in Chapter 2.3. Other models include the Paragraph Vector Topic Model (Lenz & Winker, 2020), which uses doc2vec (Le & Mikolov, 2014) to compute vector representations of documents and Gaussian mixture models to cluster the document vectors into topics. Similarly, top2vec (Angelov, 2020) uses doc2vec or BERT (Devlin, et al., 2018) sentence transformers to create document and word vector representations, and it clusters documents using HDBSCAN (Campello, et al., 2013). The Embedded Topic Model (Dieng, et al., 2020) is a generative probabilistic model where each document is represented as a mixture of topics, each word is represented by an embedding and each topic is a point in the embedding space. Lisena et al. (2020) compare and review other additional models, including modern models based on neural networks, and describe metrics for their evaluation.

While this project only makes use of topic models and textual information, there are also various hybrid methods that combine text and citation information in order to produce publication clusters and to identify topics in scientific fields (Boyack & Klavans, 2020; Janssens, et al., 2008; Liu, et al., 2010; Yu, et al., 2017). In general, experience suggests that hybrid approaches produce more accurate clusters than text-only or citation-only approaches. However, there is no ground truth of the "correct" scientific field available, and no one method can be declared superior to others, and therefore the choice of method for clustering publications depends on the objectives of the analysis (Suominen & Toivanen, 2016). Furthermore, this project

intends to identify research topics among Finnish publications, not to propose a new classification scheme.

In addition to identifying research topics, this project aims to find related research collaborations or research ecosystems. LDA and other traditional topic models do not usually describe how research communities or research ecosystems contribute to a certain topic. Research communities are typically detected using co-authorship networks, where the nodes contain only information about authors: consequently, no information is obtained on how a community relates to topics (Yan, et al., 2012). Yan et al. (2012) combined a co-authorship network and a topic model approach to study the dynamic interactions of topics and communities.

Suominen and Toivanen (2016) present a map of Finnish science based on 144,081 publications between 1995 and 2011. They use the Latent Dirichlet Allocation topic modelling method to cluster Finnish science into 60 topics. In addition, they merge the topic assignments of each publication with the OECD major classification of each publication to create a network visualisation. Employing a modularity network clustering algorithm, Suominen and Toivanen identify five thematic communities: medical research, biological research, chemical and physical sciences, Earth science, and a community containing information and communication technology-related science and social sciences, as visually illustrated in the Map of Science (Hajikhani & Suominen, 2018).

Following these previous studies on topic models and the mapping of Finnish science, this report enhances topic modelling methods by using word embeddings to capture the semantic knowledge at document level. To describe research collaborations between organisations, co-authorship networks are formed based on the set of publications in each topic. Throughout this report, the term “research ecosystem” refers to a group of organisations co-publishing on a topic.

2. Data and methods

The analysis workflow consists of data acquisition, data preprocessing, topic modelling, the building of collaboration networks and computing citation impact. First, data is retrieved from the Web of Science (WoS) Core Collection in XML format. The abstracts, titles and keywords of publications are then preprocessed. Next, selected publications are classified into research topics using unsupervised topic modelling methods. Collaboration in the identified topics is studied by forming co-authorship networks between organisations. Finally, the citation impact of co-authored publications is evaluated by computing the top 10 indices by topic.

2.1. Dataset

In total 106,736 publications (articles, reviews, letters, books) with at least one author with a Finnish affiliation, written in English and published in 2008–2019, were retrieved from the Web of Science Core Collection. We chose to limit the publication years to the period 2008–2019 for several reasons. We were interested in finding relatively recent, still active topics, and did not want to look too far into the past. On the other hand, topic modelling algorithms require large datasets in order to provide meaningful results. In addition, a longer time period allows us to identify and analyse topics with relatively small numbers of publications.

The number of publications from 2008 to 2019 is shown in Figure 1 and the publication types are summarized in Table 1. Figure 1 shows that there has been a steady rise in the number of publications in 2008–2019, which is explained by the increasing number of journals in the database, changes in publication patterns in different scientific disciplines and countries, and changes in the science system in general. One of the main limitations of the Web of Science database is that it has a low coverage especially in the humanities and social sciences, compared to the national VIRTA publication information service. Weaknesses in the data and methods are discussed in more detail in Chapter 4.

Table 1. Publication types in the dataset

Publication type	Publication count
Article	100929
Review	5533
Letter	185
Book	89

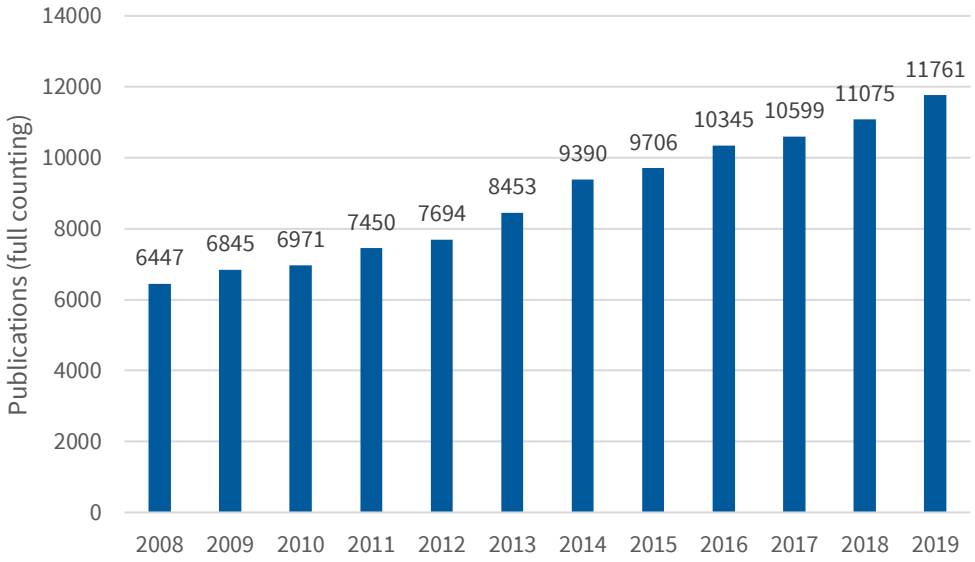


Figure 1. Selected publications in 2008–2019

2.2. Data preprocessing

After raw data acquisition, the data were preprocessed using Python 3.7. Only the titles, abstracts and keywords of publications were considered for further topic model analysis; other metadata such as author and journal names as well as citations were excluded.

During data preprocessing, all words were tokenised, converted to lowercase, and numbers and special characters except hyphens ("-") were removed. Stopwords with little semantic value, such as prepositions and articles, were detected and removed using the list of predefined stopwords in English in the nltk package. To enrich the dictionary of words, unigrams and bigrams were formed: unigrams "*machine*" and "*learning*" form the bigram "*machine_learning*". Too frequent words are not very discriminant, and for too rare words it is hard to capture the contextual information. Therefore both unique words and words that appear in more than 90% of the publications may act as noise in topic modelling and were removed in the preprocessing step. Without this step, nearly all topics would include common words such as "study", "measurement" and "result".

Finally, data quality was controlled by checking that no duplicates or missing values remained in the sample. In addition, a random sample of 100 publications was selected and data were manually checked to ensure quality.

2.3. Topic modelling: from titles and abstracts to topics

Research topics were identified from preprocessed data topics using unsupervised topic modelling methods. The selected latent feature LDA model (Nguyen, et al., 2015) aims to improve the standard LDA model (Blei, et al., 2003) by using word vector representations. In the model, each publication *d* is represented as a probability distribution over latent topics ($\theta_d \sim$ Dirichlet(α)), and each topic *z* is represented as a probability distribution over words ($\phi_z \sim$ Dirichlet(β)). The original LDA Dirichlet multinomial component that generates words from topics is extended with two matrices of latent feature vectors τ and ω associated with topics and words, respectively. In the implementation used, the word and document vector presentations were trained simultaneously by using the distributed bag-of-words variant of doc2vec (Le & Mikolov, 2014) in the gensim package.

To generate a publication d, a distribution over topics θ_d is first drawn from a Dirichlet with prior parameter α. Next, for each word w_d, a topic $z_d \sim$ Categorical (θ_d) is sampled from the topic distribution in the publication. Then, the model uses a binary variable $s_d \sim$ Bernoulli (λ) to choose whether the word is generated by the Dirichlet multinomial component or the latent feature component. The posterior distributions of the latent variables θ_d, ϕ_z and z, i.e., which topics are significant in each publication and which words are important in these topics, are inferred using collapsed Gibbs sampling. For further details about the algorithm, the inference process and how to learn latent feature vectors, see Nguyen et al. (2015).

The model requires that four key hyperparameters be set in advance: the number of topics, two hyperparameters α and β that control prior assumptions about the topic distribution of publications and the word distribution of each topic, and λ that is the probability of a word being generated by the latent feature model. In general, a low α value results in the publication having only a few dominant topics. In this project, each publication was finally assigned only to the most significant topic. Similarly, a low β value results in each topic consisting of only a few dominant words, which is often useful for the interpretability of topics (Griffiths & Steyvers, 2004). Changing these values highly affects the resulting topics. Multiple versions of the model were trained to evaluate the optimal hyperparameter settings. For example, the number of topics was scanned from 5 to 2,500. The final model settings were selected based on the various coherence metrics and human judgement on interpretability and distinguishability from other topics: in particular, the number of topics was set to 1,026. For more details about the parameters and evaluation, see Appendix G.

2.4. Measuring research impact

To evaluate scientific impact in the research topics identified, the top 10 citation index was calculated for each topic. The top 10 index is the proportion of

publications in the focal set belonging to the most frequently cited 10% in their respective fields when compared to publications of equal age. With proper normalisation, the world average of the top 10 index will be 1.0 for each field, each year (Waltman & Schreiber, 2013). For example, a top 10 index value of 1.2 means that 12% of the publications in the focal set are among the most cited 10%. Similarly, an index value of 0.8 means that only 8% of publications are among the most highly cited 10% (in other words, 20% below the world average).

For the calculation of citation impact, a fractionalised counting method at the level of fields, countries and organisations was used instead of full counting. In full counting, a weight of one is assigned to each publication. In fractionalised counting, the weight of each publication is divided between the collaborating organisations or countries, as well as the scientific fields of the publication. For example, if a publication is co-authored by two organisations, each organisation will receive a weight of 0.5 in the calculation of the top 10 index. Fractionalised counting provides results that are properly field normalized or scaled and therefore allows for comparisons (Waltman & van Eck, 2015). By contrast, any impact indicator based on full counting would easily result in meaningless results because of the absence of a proper scale.

The number of citations varies widely across scientific fields. Also, more recent publications have had less time to accrue citations than older ones. Therefore, the number of citations received by a publication is always compared to the number of citations received by publications of the same age in the same field. The field of a publication is determined by the Web of Science subject classification. Calculating the top 10 citation index is not meaningful if the number of publications is too low. In this report, it is required that the number of fractionalised publications in a topic is greater than 40. The citation window is open, i.e., all citations are taken into account. Because publications from 2018 and 2019 have not yet had sufficient time to mature for impact analysis, they are included in publication volumes but excluded from the calculation of the top 10 indices.

It should be noted that our aim was to find research ecosystems that produce research with high citation impact, not to assess the level of Finnish research on various topics. In particular, the citation indicators do not measure research quality per se (Aksnes, et al., 2019).

3. Results

The topic model generated a total of 1,026 research topics. The number of publications in each topic varies between 18 and 862, with the mean being 104.0 and

median 81.5. Most of the generated topics represent easily interpretable, large-scale research phenomena. Topics with the most publications represent broad disciplines from education and business management to wireless networks, renewable energy and gut microbiota, whereas the smallest topics tend to be very specific such as distinct animal or plant species (e.g., frogs, spruces) or diseases (e.g., carpal tunnel syndrome). In addition, a few small low-quality topics were detected in which no clear connections can be found between more than a couple of words or publications.

In many topics the same general theme can be covered from different perspectives. For example, there are many topics related to oncology, usually distinguished by an organ of interest ("prostate cancer", "cervical cancer"). Therefore, it is not possible to directly deduce which are the largest overall themes. Similar topics could be merged together algorithmically by using cosine similarity or Hellinger distance, for example. However, algorithmic merging is by no means definitive or unambiguous, and reliable classification would require subject matter expertise. On the other hand, some topics include publications that are somewhat loosely connected to each other: for example, the topic "Arctic maritime" covers a wide range of research from ships to port waste management, from navigation systems to maritime safety and to ice deformation in Arctic waters.

It should be noted that not all publications in a given topic are necessarily strictly related to that topic, and not every publication concerning the topic in question has been clustered into it. Many significant topics and research ecosystems may remain unrecognized due to data and methods limitations. In particular, research related to the humanities and social sciences may be underrepresented due to low coverage in the Web of Science database. The limitations and weaknesses in this project are discussed in more detail in Chapter 4.

Because of the large number of topics, the focus henceforth is on those topics with high citation impact. Because these high-impact topics are mainly related to engineering, technology and natural sciences, some organisations may appear more often than others due to their research profiles. In this section, topics representing various scientific fields are shown as examples, showcasing their variety. All topics with high citation impact are shown in more detail in Appendix A.

3.1. Topics with high citation impact

In order to find high-impact research, top 10 indices were calculated for publications by topic. This yielded 65 topics with high citation impact (top 10 index over 2.0), 10 of which had very high citation impact (top 10 index over 3.5). These 10 topics are shown in Table 2; the rest of the high-impact topics are shown in Appendix A. In order

to gain a clearer overview, each topic is human-labelled[2] by selecting the most descriptive keywords among the output keywords of the topic model.

Table 2. Topics with the highest citation impact as indicated by the top 10 index together with full and fractionalised count of publications. The first column, topic number, serves as a label only.

Topic	Keywords	Publications (full counting)	Publications (fractionalised counting)	Top 10 index
620	permanent-magnet synchronous motors, sensorless control	57	47.3	5.33
220	computer vision, local binary patterns, image classification	104	66.1	4.69
215	denoising, filtering	116	76.4	4.13
314	unmanned aerial vehicles, remote sensing	60	42.7	4.08
104	MIMO systems	158	108.6	4.01
81	millimeter wave MIMO systems	156	99.6	3.78
38	gamification, user acceptance and adoption in social media and virtual worlds	207	151.9	3.74
480	video coding and compression	73	47.7	3.72
141	RFID technologies	143	95.2	3.61
92	fading, fading channels	174	105.1	3.51

Most of the topics with the highest top 10 indices are related to various aspects of computer science, electrical and electronic engineering, telecommunications, environmental sciences, materials science, as well as business and management. The proportions of the most dominant WoS subject categories in the topics with the highest top 10 indices are shown in Table 3. The full list of the WoS subjects in the high-impact topics is shown in Appendix C.

[2] It should be noted that the topics are labelled by an analyst without subject matter expertise. Keywords that describe topics in more detail are listed in Appendix E.

Table 3. Proportions of the most dominant WoS subject categories in the topics with the highest citation impact

Topic	Keywords	WoS subjects	[%]
620	permanent-magnet synchronous motors, sensorless control	Engineering, Electrical & Electronic	43.3
		Engineering, Multidisciplinary	15.8
		Automation & Control Systems	12.0
		Instruments & Instrumentation	9.0
220	computer vision, local binary patterns, image classification	Computer Science, Artificial Intelligence	40.6
		Engineering, Electrical & Electronic	24.7
215	denoising, filtering	Engineering, Electrical & Electronic	41.8
		Computer Science, Artificial Intelligence	17.0
		Computer Science, Software Engineering	5.3
314	unmanned aerial vehicles, remote sensing	Remote Sensing	29.1
		Engineering, Electrical & Electronic	12.3
		Imaging Science & Photographic Technology	10.1
		Telecommunications	5.9
		Geosciences, Multidisciplinary	5.6
104	MIMO systems	Engineering, Electrical & Electronic	50.1
		Telecommunications	37.4
81	millimeter wave MIMO systems	Engineering, Electrical & Electronic	44.6
		Telecommunications	37.8
38	gamification, user acceptance, adoption in social media & virtual worlds	Information Science and Library Science	21.4
		Computer Science, Information Systems	8.8
		Psychology, Multidisciplinary	8.7
		Psychology, Experimental	7.8
		Communication	7.5
		Business	6.6
480	video coding and compression	Engineering, Electrical & Electronic	52.3
		Computer Science, Information Systems	12.8
		Telecommunications	11.1
		Computer Science, Software Engineering	10.5
		Computer Science, Theory & Methods	5.9
141	RFID technologies	Engineering, Electrical & Electronic	41.9
		Telecommunications	23.6
		Instruments & Instrumentation	6.1
92	fading, fading channels	Telecommunications	49.5
		Engineering, Electrical & Electronic	37.1

The same broad topic can include publications from many different WoS subjects. For example, topic 38 is related to gamification and user acceptance and adoption in social media and virtual worlds, and its publications have a variety of WoS subject categories from Information Science and Library Science to Computer Science and Psychology as well as Business and Management. Similarly, a topic related to unmanned aerial vehicles and remote sensing (topic 314) has publications from various WoS categories including Remote Sensing, Electrical & Electronic Engineering, Imaging Science & Photographic Technology, Physical Geography as well as Telecommunications and Forestry. On the other hand, other topics such as 220 related to computer vision, especially in local binary patterns and image

classification, have only a couple of dominant WoS subjects, in this case Computer Science & Artificial Intelligence and Electrical & Electronic Engineering.

Figure 2 shows an example of a co-authorship network, where the nodes are organisations. The organisations involved therefore form a research ecosystem, i.e., a group that co-authors on a topic. Some of the names of the organisations are abbreviated; for the full names, see Appendix F. The size of the circle indicates the whole count of publications. The sizes are scaled for each figure separately and cannot therefore be compared between topics. The colour of the circle describes the citation impact with a colour scale on the right. Red indicates that the impact of an organisation in a topic is below average; green that itsimpact is around average (i.e., the top 10 index is around 1.0); blue that the impact is above average; and black that the impact is exceptionally high. Because the topics are small, the impact of individual organisations is expressed only in colours. For the same reason, the top 10 indices of the topics can be higher than those of scientific fields.

Only the largest organisations in each topic are shown as nodes, while smaller ones are merged into “Others”. Similarly, all foreign organisations are grouped together as “Foreign” and companies as “Industry“. The edges between the nodes describe the whole count of co-authored publications. The thickness of an edge indicates the number of publications, and the colour describes the citation impact. However, if the number of publications is 0–2, no edge is drawn, and if there are only a few co-authored publications, the edge is a thin black line.

Key figures are shown on the top left of the figure: N is the number of publications (full counting), τ_{10} is the top 10 index, θ_{Int} is the proportion of international co-authored publications and θ_{Dom} is the proportion of national co-authored publications. In the remaining proportion of publications, all authors work in the same organisation in Finland. All proportions are based on whole counts. The exact numbers may differ slightly between the summary tables and figures of research ecosystems because the latter are based only on publications published in 2008–2017. All research co-authorship networks in the high-impact topics can be found in Appendix B.

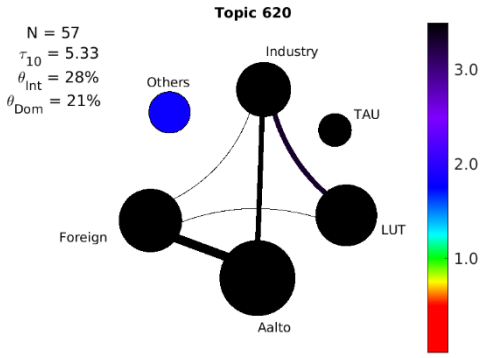


Figure 2. permanent-magnet synchronous motors, sensorless control

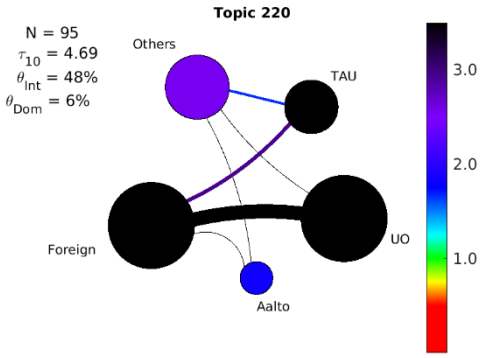


Figure 3. computer vision, local binary patterns, image classification

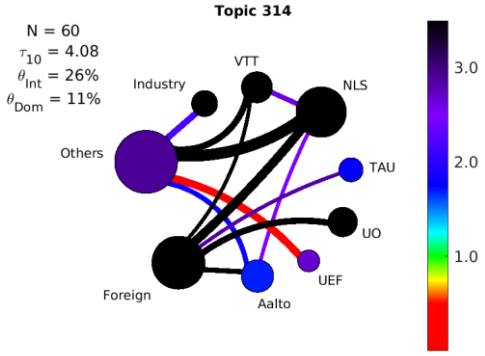


Figure 4. unmanned aerial vehicles, remote sensing

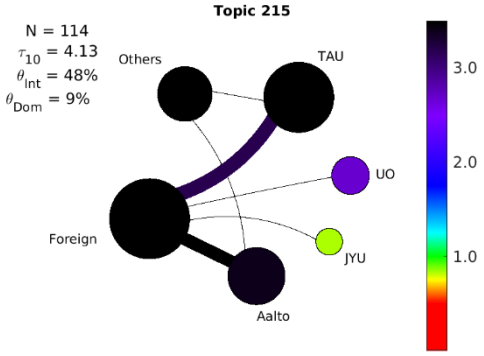


Figure 5. denoising, filtering

Above: Examples of co-authorship networks. N is the number of publications, τ_{10} is the top 10 index, θ_{Int} is the proportion of international co-authored publications and θ_{Dom} is the proportion of national co-authored publications. The colours describe the citation impact. For more keywords that describe topics, see Appendix E. For the full names of the organisations, see Appendix F.

For example, Figure 2 shows that especially Aalto University, Lappeenranta-Lahti University of Technology (LUT) and Tampere University conduct very high-impact research in topic 620 related to permanent magnet motors, but they do not co-author publications together. Aalto University co-authors mostly with foreign organisations, and Aalto University and LUT also co-author with industry. However, given the small number of publications, caution must be applied when interpreting the results.

Figure 6 shows the shares of national and international co-authored publications in the selected high-impact topics. For all topics, see Appendix D. In general, there is more international than national collaboration in the identified high-impact topics, signalling that international co-authored publications have higher citation impact. However, co-authoring patterns seem to depend on the topic. In some topics there is an emphasis towards international collaboration, while national collaboration is almost an exception. Such topics include many related to telecommunications and electrical engineering, such as topic 92 (fading and fading channels) and topic 81 (millimeter wave MIMO systems).

In many other topics, such as topic 10 (laser scanning, point clouds, Light Detection and Ranging; especially in forestry), there is more national than international collaboration, and national collaboration has mostly high citation impact, as shown in Figure 7 and in Appendix D. Especially in topics related to life sciences, it is common to have many active organisations and collaborations, as illustrated in Figure 9 and Figure 10.

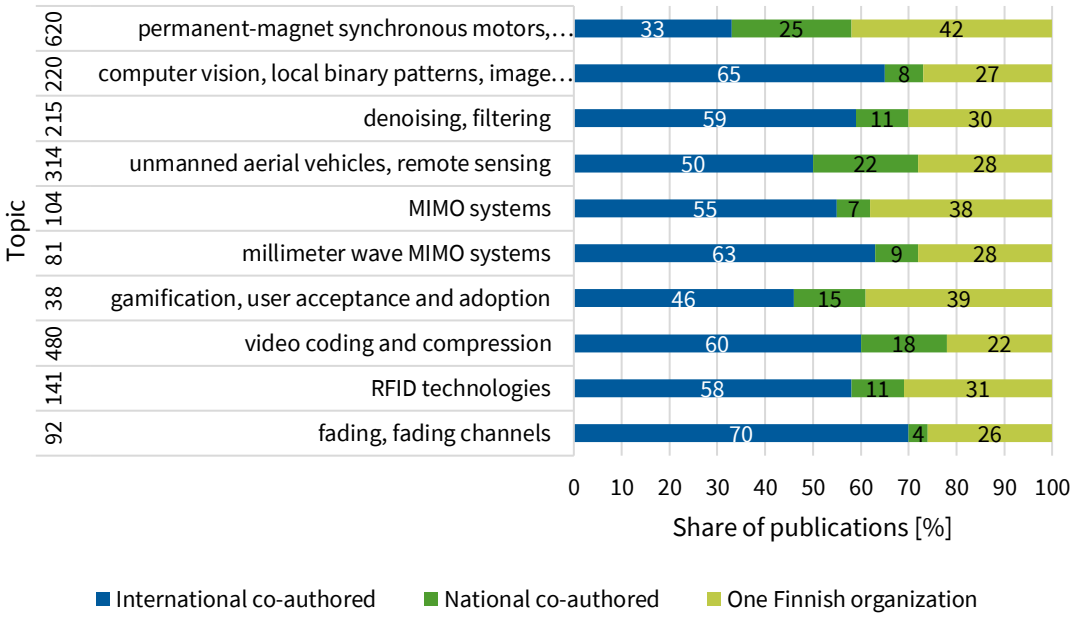


Figure 6. Shares of national and international co-authored publications in the selected topics

Citation impact can also be strongly polarised within topics: among all the organisations active in a topic, it is possible that only a few are producing high-impact research. Therefore, a high national citation impact does not directly imply that research in Finland is high-impact across the board (for example, as shown for topic 300 in Figure 8). However, the same organisation can conduct high-impact research on a single topic related to a broader theme, but be weaker on another

related topic. If an organisation is found to produce lower-impact research, that may be due to various reasons, such as misclassified publications.

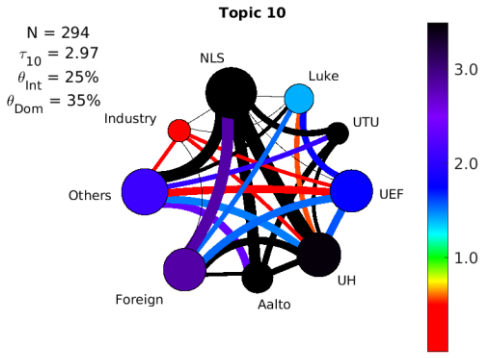


Figure 7. laser scanning, point clouds, LiDAR*

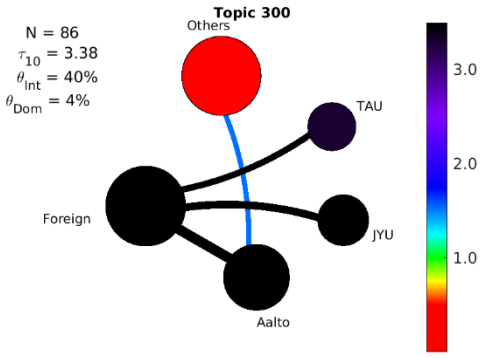


Figure 8. particle swarm optimization*

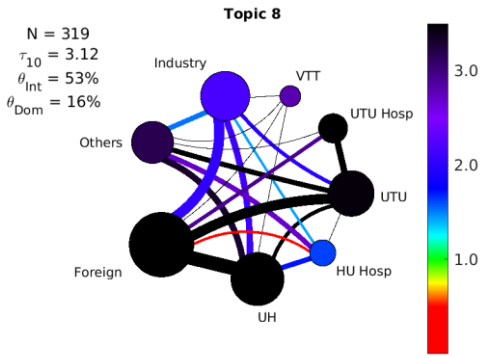


Figure 9. gut microbiota*

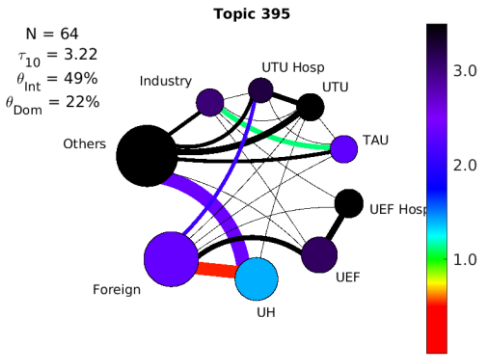


Figure 10. DNA methylation*

3.2. Topics related to the Academy of Finland's Flagship Programme

The Academy of Finland Flagship Programme is intended to promote and support high-quality research. After the third call, the programme now comprises 10 Flagships (Academy of Finland, 2021). Each of the flagship themes also emerged in

* For more keywords that describe topics, see Appendix E. For the full names of the organisations, see Appendix F. N is the number of publications, τ_{10} is the top 10 index, θ_{Int} is the proportion of international co-authored publications and θ_{Dom} is the proportion of national co-authored publications. The colours describe the citation impact.

our data-driven topic modelling analysis. However, the project did not determine the topics of each Flagship's individual publications.
For example, the 6G Flagship hosted by the University of Oulu focuses on developing wireless technology and mobile communication. Topic 1 is directly related to such themes: the topic keywords detected include "*wireless networks edge_computing iot ad_hoc wireless_sensor virtualization communications mobile_networks privacy secure*". In addition, several topics related to various technologies for wireless communications, such as MIMO (multiple-input, multiple-output) were identified among the topics with the highest top 10 indices. Examples of three identified topics linked with the 6G Flagship themes are shown in Figure 11, Figure 12 and Figure 13.

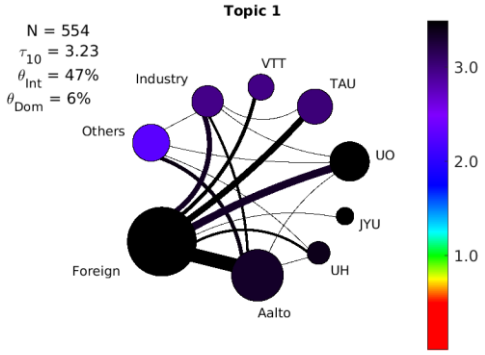


Figure 11. wireless networks*

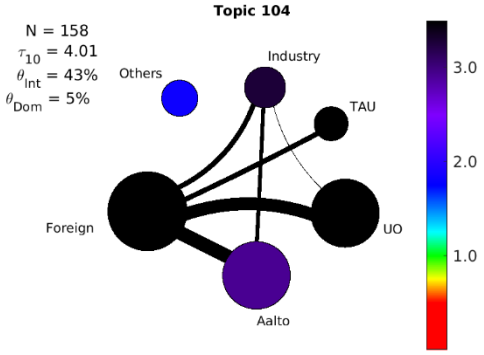


Figure 12. MIMO systems*

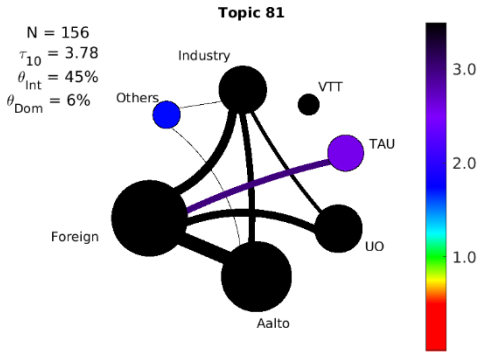


Figure 13. millimeter wave MIMO systems*

The Finnish Centre for Artificial Intelligence (FCAI), hosted by Aalto University, the University of Helsinki and the VTT Technical Research Centre of Finland, contributes to research on artificial intelligence and machine learning. For example, topic 15 (Figure 14) directly addresses machine learning, as indicated by the topic keywords

"*support_vector classifiers machines classifier supervised kernel clustering feature discriminant unsupervised classification neural_networks*". In addition, there are several other topics with applications of machine learning in various fields.

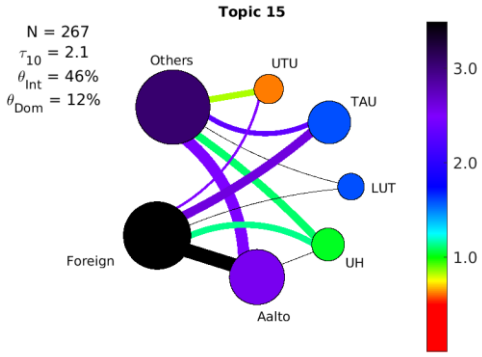


Figure 14. machine learning, classification*

3.3. Topics related to climate change

Climate change is one of the key themes in the current international debate, and resolving its challenges through research and innovation is one of the main targets of Finland's RDI roadmap (2020).

Climate change has been addressed in several topics. Two examples of research ecosystems in topics related to climate change are shown in Figure 15 and Figure 16. In topic 88, publications cover the effects of climate change on biodiversity and species, and in topic 170 publications describe the planning and monitoring of conservation and protected areas. As mentioned earlier, all publications related to climate change may not have been classified into these topics. For the full names of the organisations, see Appendix F.

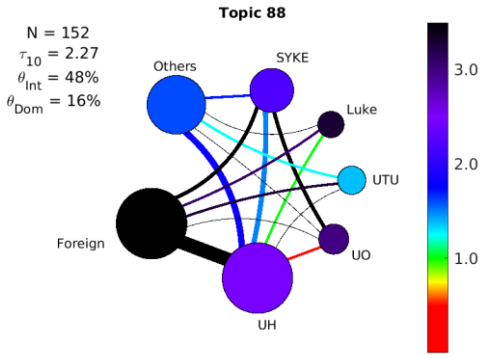


Figure 15. climate change, biodiversity, species*

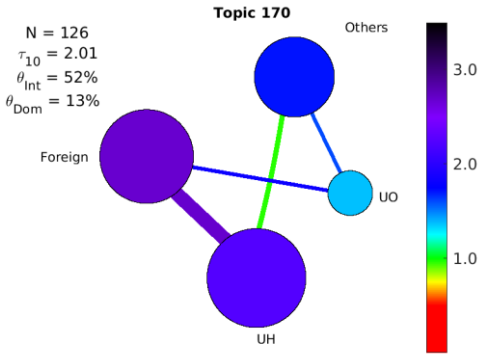


Figure 16. biodiversity, conservation, protected areas*

3.4. Other examples of topics

To further illustrate the variety of topics with high citation impact, this section highlights a few more example topics and related ecosystems.

There are several topics with high-impact research in the field of medical and health sciences. Topic 449 (Figure 17) is related to bariatric surgery, obesity and weight loss, while topic 195 (Figure 18) concerns liver diseases. Topic 299 (Figure 19) covers research in dementia and Alzheimer's disease. The role and mechanisms of microRNA are addressed in topic 242 (Figure 20). As can be seen in the figures, there is a lot of active and mostly high-impact collaboration between universities and hospitals. In these figures, SHP refers to hospitals within the hospital districts, excluding university hospitals. For other abbreviations, see Appendix F.

The topic modelling algorithm did not generate a single, general topic related to cancer and oncology. Instead, there are multiple related topics, usually distinguished by an organ of interest ("prostate cancer", "cervical cancer"), or by possible types of diagnostics and treatment. For example, some publications in topic 242 about microRNA address the role of microRNA in cancer.

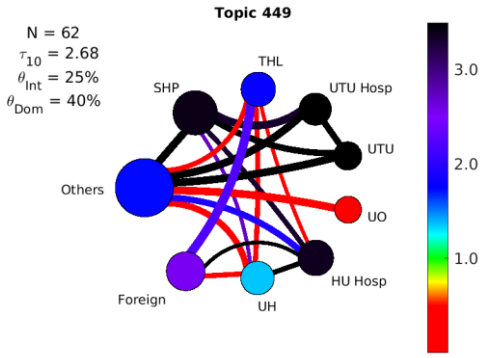


Figure 17. obesity, weight loss, bariatric surgery*

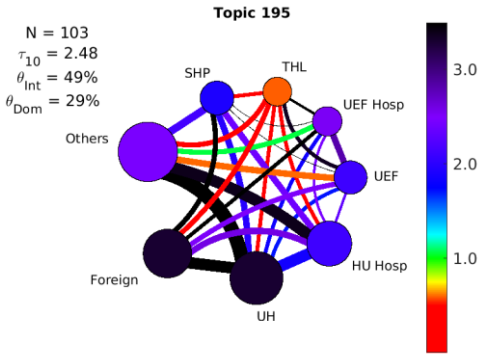


Figure 18. liver diseases*

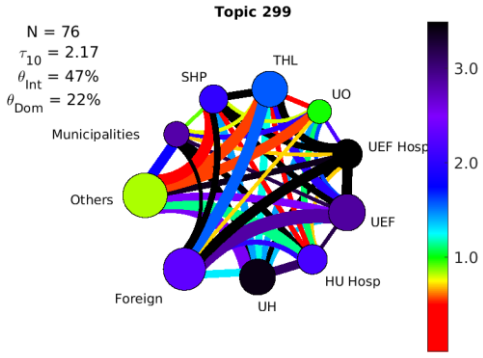


Figure 19. dementia, Alzheimers*

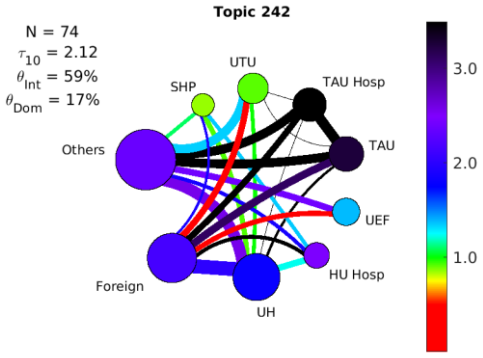


Figure 20. microRNA*

Other examples of various topics are presented in Figure 21 and Figure 22, which show two co-authorship networks related to business and management. Topic 146 addresses research about business models and servitisation, and the related topic 14 covers more publications about service and customer value creation. However, both topics share similar model keywords, including "*servitization, value_creation, customer, service_systems, dominant_logic, offerings*".

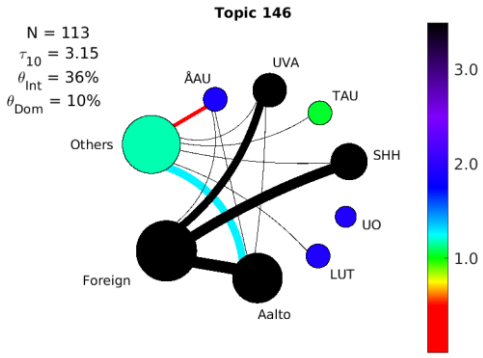


Figure 21. servitization, business models*

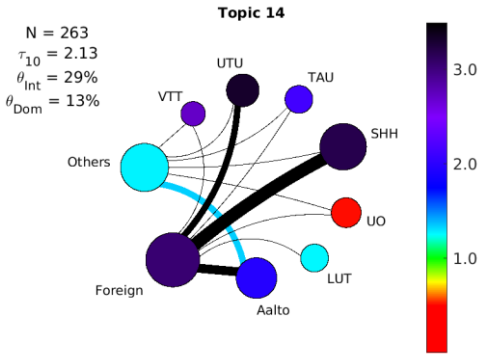


Figure 22. service and customer value creation*

Topic 217 (Figure 23) about "music" covers a wide range of research from emotions (psychology) to brain responses (neurosciences) and musicology. Another partly art-related topic with high citation impact is topic 342 about creativity. This topic contains publications not only on art making and learning, but also on how to create a work environment that nurtures creativity. The publications are thus from various research fields such as educational research, management, art and information science. In general, there is only little collaboration between organisations in this topic, as can be seen in Figure 24.

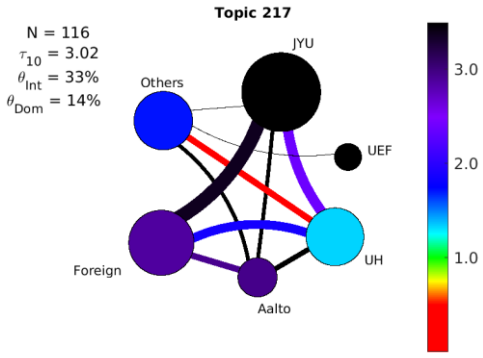


Figure 23. music*

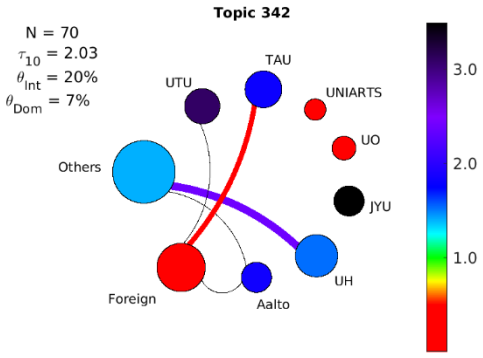


Figure 24. creativity*

As for agriculture, there are two topics with high citation impact related to veterinary sciences and dairy & animal sciences. Topic 357 contains publications about pigs and sows, and topic 442 contains publications about dairy cattle. The two main national organisations in both are the University of Helsinki and the Natural Resources Institute Finland (Luke), as illustrated in Figure 25 and Figure 26.

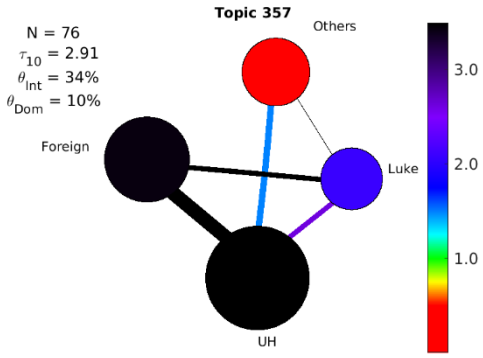


Figure 25. pigs, sows*

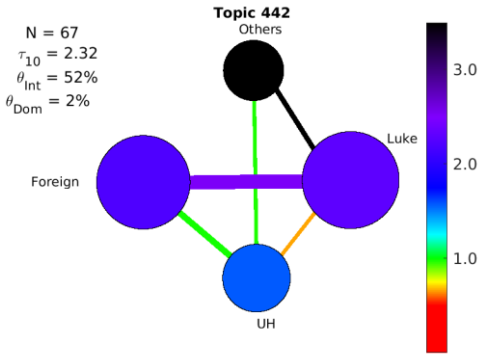


Figure 26. dairy cattle*

3.5. Dynamics of topics

In addition to the previously presented analyses of topics with high citation impact, we studied the dynamic changes of all generated topics. A linear regression analysis was conducted to study trends in publication volumes over time. As the total number of publications increases from year to year in 2008–2019, the number of publications for the topics has been normalised with respect to the annual total for each year. Examples of topics with the highest increasing shares of publications are presented in Figure 27. The figure shows that the number of publications in renewable energy, wireless networks, and education research, for example, has risen steadily in recent years. The keywords for the topics include the following:

- topic “education”: “*teacher student teaching pedagogical education learning classroom inquiry pedagogy educational*”
- topic “renewable energy and energy systems”: “*energy_system electricity energy_storage wind_power side_management renewable power_systems combined_heat solar_power energy_sources*”
- topic “wireless networks”: “*wireless networks edge_computing iot ad_hoc wireless_sensor virtualization communications mobile_networks privacy secure*"
- topic “big data analytics”: “*analytics big_data information data_mining cloud_computing customization technologies computing iot internet*”
- topic “urban planning”: “*urban_planning cities urban infrastructure metropolitan governance sustainability urban_areas participatory gis*”
- topic “metallurgy”: “*strength_steel weld high_strength joints steel laser_welding fatigue crack toughness stainless*”.

No particularly growing or peaking topics were found. Furthermore, none of the identified topics showed a rapid decrease in the number of publications in 2008–2019.

However, the figure does not give a proper overall picture of possible research trends, as some of the research themes are covered in several different topics. Moreover, although the generated topics represent active topics in the 2010s, some of them may have been studied for a long time. Therefore, the topics found may not represent novel or emerging research areas. It should also be noted that in some topics the number of publications is small and the changes may reflect random variation and statistical fluctuation rather than a real trend. For the same reason, it is not meaningful in this context to study temporal changes in citation impact. The increasing trend can also be partly explained by changes in the database: it is possible that WoS has recently indexed more publications related to these topics.

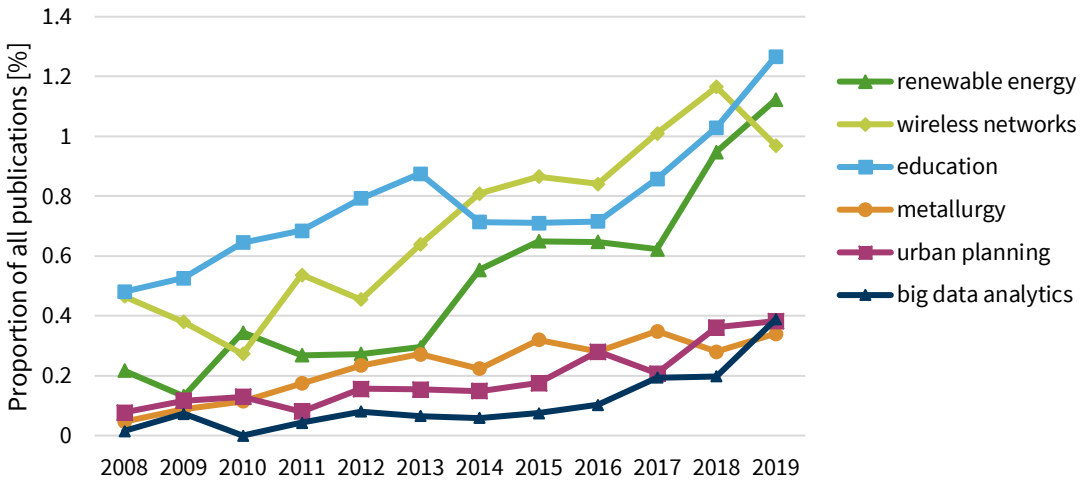


Figure 27. Relative publication volume in selected topics in 2008–2019 normalised to yearly count of publications to account for overall increase in publication volume. Note that the figure does not give a proper overall picture of possible research trends, as some of the research themes are covered in several different topics. See the text for more discussion and for keywords describing topics.

4. Limitations and future directions

Although the purpose of this exploratory project was to identify research topics with high citation impact and related research ecosystems, even significant topics and ecosystems may have remained unrecognized due to data and methods limitations.

Furthermore, it should be noted that not all publications in a given topic are necessarily strictly related to that topic, and not every publication concerning the topic in question has been clustered into it. While the methods employed are suitable for mapping research topics at scale, they cannot be used for accurate classifications or for determining the exact size of research areas. Therefore, the results of the analysis are not useful for decision-making purposes but should be viewed as a starting point for more in-depth analyses rather than a suggestion for a new classification scheme.

4.1. Dataset

The selected data sample is only a subset of all Finnish research published in 2008–2019. The coverage of Finnish publications in the Web of Science database varies greatly by discipline and organisation (e.g., Leino, 2020). For example, in humanities the coverage is only 21.3% and in social sciences 58.6% when compared to the national VIRTA publication information service, whereas coverage figures in natural sciences and medical sciences are higher (Auranen & Leino, 2019). Another strong restriction comes from the decision to focus on only specific publication types (articles, reviews, letters, books): in particular, publications in the field of technology, such as computer science, are usually published as Proceedings Papers and Meeting abstracts (Freyne, et al., 2010) and therefore may now be underrepresented in the analysis. Furthermore, the project was limited to analysing publications written in English. However, especially in humanities and social sciences it is also common to publish in Finnish and Swedish (Pölönen & Auranen, 2020). Given these issues with low coverage, organisations' publication activities should not be evaluated based solely on the WoS dataset (Leino, 2020).

In some topics the number of publications is very small, which means that statistical uncertainties are high and that the set of publications is not necessarily suitable for a reliable bibliometric analysis. In addition, the full publication counts should not be summed across organisations or fields of science.

4.2. Research ecosystems

In this report, a research ecosystem is loosely defined as any group of organisations that co-author publications in a topic. Therefore it does not necessarily represent any officially formed or funded ecosystem. Furthermore, the co-authors are recognized only at the organization level, i.e., not at the level of research units, departments, groups or individual researchers.

Some organisations receive strong emphasis in the analysis due to the nature of the often rather technical topics identified. However, other organisations may also

produce high-impact research in these topics. Multiple research groups within the same organisation may also be involved in the same topic. In addition, some topics are broad and may include publications from various fields. This does not mean that organisations necessarily do interdisciplinary research between those fields: collaboration can still happen strictly within one discipline.

Bibliometric datasets do not provide fully accurate means for a more in-depth understanding of research ecosystems. The analysis here focuses only on co-authorships and neglects joint projects between organisations, for example. Other sources could be used to assess the size, growth and economic impact of the ecosystems. Measures of size might include the number of researchers at different career stages, full-time equivalent (FTE), budget and funding (e.g., via the education statistics portal Vipunen administered by the Ministry of Education and Culture). Economic impact could be studied based on patents and additional research, development and innovation (RDI) indicators.

4.3. Top 10 index as an indicator of impact

The top 10 citation index reflects only one side of scientific impact and should not be used on its own to evaluate research. In particular, citation indicators do not alone reflect the quality of research (Aksnes, et al., 2019). It is important to note that citation analyses are affected by statistical time delay and only reflect history. In this report, impact indicators have been calculated for publications published in 2008–2017. More recent publications have not had enough time to accrue citations and therefore the most recent, emerging high-impact research topics and ecosystems may remain unrecognized. Furthermore, citation-based indicators are not stable in time because the number of citations increases over time.

Some publications may have been misclassified to a topic to which they do not belong. If a wrongly-classified publication does not belong to the most cited 10% in its own field, it will falsely decrease the top 10 index of the topic by order of 0.01. If it does belong to the most cited publications, it will slightly increase the top 10 index. The effect may be amplified if the number of publications in a topic is small.

4.4. Topic modelling method

Topic models have been shown to provide a way of identifying topics of scholarly publications automatically at scale. However, the data-driven topic modelling method used here is unsupervised and does not take into account the field of the publication, citations, or any ontology-based classification. In addition, it requires multiple parameters to be set. Changing the (hyper)parameters of the method can lead to major differences in the quality and content of topics, and therefore the

method cannot identify topics unambiguously. It was found that optimising a single coherence metric alone does not guarantee satisfactory results. Instead, several coherence metrics were combined to guide human judgement on the number of topics and other final hyperparameters. If there are only a few topics, they reflect only large fields such as physics and engineering, while smaller topics of local interest such as those closely related to humanities or social sciences will not arise. On the other hand, having too many topics leads to very specific topics with only a handful of publications, or duplicates of similar topics. It should also be noted that the role of the algorithm is to find a predetermined number of topics, regardless of whether they actually exist in the data.

The method only accepts titles, abstracts and keywords as input. Some publications have only a very short abstract or title and zero or only a few keywords, which can adversely affect the quality of the results. Furthermore, during preprocessing of the text, even important words can be removed because of their rarity or frequency. A large imbalance between the number of publications in different scientific fields has an effect on the input vocabulary.

Topics could be combined to form broader themes according to their similarity, for example by using cosine similarity or Hellinger distance between topics. This way it is possible to cluster different types of cancer, for instance, into one broad cancer topic. However, this may give rise to overly generic themes and problems with continuity: if swamps are related to wetlands, and wetlands are related to water, water is related to water quality, and water quality to sewage treatment, are swamps and sewage treatment still connected? Combining topics reliably requires subject matter expertise.

Manual inspection may reveal publications in a topic that do not necessarily belong there. Filtering out such publications is not trivial, and we attempted to do so among other things by removing wrongly-classified publications based on their WoS subject and low probability scores. Another approach would be to take account of citation information. However, neither of these approaches are unproblematic when the topic is truly multidisciplinary in nature. The proper filtering of publications would also require subject matter expertise.

4.5. Future directions

In order to obtain a more comprehensive picture of Finnish research, future analyses should consider including other publication types and publications written in other languages (Finnish, Swedish). Instead of the Web of Science, another option is to use publication data collected through the national VIRTA publication information

service. That would make it possible to identify a wider range of topics, especially those related to humanities and social sciences.

As for topic modelling, future work could extend this study by making use of large, pretrained neural network models such as SPECTER (Cohan, et al., 2020), which also uses citation information to relate documents. However, while such transformer-based models, including BERT, have gained popularity, they may require careful fine-tuning before being applicable to this kind of tasks. Further studies might involve systematically reviewing different models and taking a closer look at hybrid models that combine text analytics and citation information.

It might also be useful to compare topics and ecosystems in Finland with other countries. However, topic modelling of the entire WoS database in 2008–2019 would have yielded thousands of different topics, among which smaller or Finland-specific topics might not have clearly stood out. Further studies might also explore the dynamics of topics and ecosystems in closer detail.

5. Conclusions

The aim of this exploratory project was to examine the potential for combining topic modelling and bibliometric methods in identifying research topics and related research collaboration or research ecosystems in Finland. To find research topics, topic modelling methods were applied to the abstracts, titles and keywords of 106,736 research papers published in 2008–2019. Research ecosystems were sought by building co-authorship networks that describe collaboration between different organisations in each generated topic. Finally, citation impact was evaluated by considering the share of highly cited publications within each topic.

Several research topics with high citation impact and related organisations were identified. However, even significant topics and ecosystems may have remained undetected due to limitations in the dataset (e.g., low coverage of publications in humanities and social sciences and) and weaknesses of the method used (e.g., changing model parameters leads to very different topics). While the methods chosen are useful for mapping research themes at scale, they are not suited for accurate classifications or for determining the exact size of research areas.

Despite its exploratory nature, this project provides some insights into research topics in Finland and demonstrates that bibliometric analyses and scientific publications can be used to find research collaboration. Further in-depth analyses that combine various data sources and more advanced methods could help to shed more light on the research landscape in Finland.

References

Academy of Finland, 2018. *State of Scientific Research in Finland 2018.* [Online]
Available at: https://www.aka.fi/en/about-us/data-and-analysis/state-of-scientific-research-in-finland/state-of-scientific-research-2018/
Academy of Finland, 2021. *Finnish Flagship Programme.* [Online]
Available at: https://www.aka.fi/en/research-funding/programmes-and-other-funding-schemes/flagship-programme/
Aksnes, D. W., Langfeldt, L. & Wouters, P., 2019. Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*, 9(1), p. 2158244019829575.
Angelov, D., 2020. Top2Vec: Distributed Representations of Topics.
Auranen, O. & Leino, Y., 2019. *Bibliometric indicator to assess the effectiveness of competitive research funding.* s.l.:24th Nordic workshop on bibliometrics and research policy.
Blei, D. M. & Lafferty, J. D., 2006. *Dynamic Topic Models.* New York, NY, USA, Association for Computing Machinery, pp. 113-120.
Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, mar, 3(null), pp. 993-1022.
Boyack, K. W. & Klavans, R., 2020. A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, 12, 1(4), pp. 1570-1585.
Campello, R. J., Moulavi, D. & Sander, J., 2013. *Density-Based Clustering Based on Hierarchical Density Estimates.* Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 160-172.
Chang, J. et al., 2009. *Reading Tea Leaves: How Humans Interpret Topic Models.* Vancouver, B.C., Canada, Curran Associates, Inc..
Cohan, A. et al., 2020. *SPECTER: Document-level Representation Learning using Citation-informed Transformers* , ACL.
Deerwester, S. et al., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), pp. 391-407.
Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
Dieng, A. B., Ruiz, F. J. & Blei, D. M., 2020. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 07, Volume 8, pp. 439-453.
Freyne, J., Coyle, L., Smyth, B. & Cunningham, P., 2010. Relative status of journal and conference publications in computer science. *Communications of the ACM,* 53(11), pp. 124-132.

Gerlach, M., Peixoto, T. P. & Altmann, E. G., 2018. A network approach to topic models. *Science Advances*, 4(7).
Glenisson, P., Glänzel, W., Janssens, F. & De Moor, B., 2005. Combining full text and bibliometric information in mapping scientific disciplines. *Inf. Process. Manage.*, 12, Volume 41, pp. 1548-1572.
Griffiths, T. L. & Steyvers, M., 2004. Finding scientific topics. *Conference Proceedings of the National Academy of Sciences*, 101(suppl 1), pp. 5228-5235.
Hajikhani, A. & Suominen, A., 2018. *The Science Map of Finland.* s.l.:s.n.
Hofmann, T., 1999. *Probabilistic Latent Semantic Indexing.* New York, NY, USA, Association for Computing Machinery, pp. 50-57.
Janssens, F., Glänzel, W. & De Moor, B., 2008. A hybrid mapping of information science. *Katholieke Universiteit Leuven, Open Access publications from Katholieke Universiteit Leuven*, 06, Volume 75.
Lau, J. H. & Baldwin, T., 2016. *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation.* s.l.:s.n.
Leino, Y., 2020. *Julkaisutietokantojen kattavuus: VIRTA vs. Web of Science.* s.l.:Bibliometriikkaseminaari 2020.
Lenz, D. & Winker, P., 2020. Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE*, 01, 15(1), pp. 1-18.
Le, Q. & Mikolov, T., 2014. *Distributed Representations of Sentences and Documents.* s.l., JMLR.org, pp. II-1188.
Lisena, P., Harrando, I., Kandakji, O. & Troncy, R., 2020. *TOMODAPI: A Topic Modeling API to Train, Use and Compare Topic Models.* Online, Association for Computational Linguistics, pp. 132-140.
Liu, X. et al., 2010. Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Journal Database. *JASIST*, 06, Volume 61, pp. 1105-1119.
Mikolov, T. et al., 2013. *Distributed Representations of Words and Phrases and their Compositionality.* s.l., Curran Associates, Inc..
Mimno, D. et al., 2011. *Optimizing Semantic Coherence in Topic Models.* Edinburgh, Scotland, UK., Association for Computational Linguistics, pp. 262-272.
Ministry of Education and Culture, 2020. *The National Roadmap for Research, Development and Innovation: Solutions for a sustainable and developing society.* [Online]
Available at: https://minedu.fi/en/rdi-roadmap/
Nguyen, D. Q., Billingsley, R., Du, L. & Johnson, M., 2015. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, Volume 3, pp. 299-313.
Paatero, P. & Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), pp. 111-126.

Pennington, J., Socher, R. & Manning, C. D., 2014. *GloVe: Global Vectors for Word Representation.* s.l., s.n., pp. 1532-1543.
Pölönen, J. & Auranen, O., 2020. *Responsible metrics for assessing competitive research funding. Case: Academy of Finland funded research.* s.l.:25th Nordic Workshop on Bibliometrics and Research Policy.
Řehůřek, R., 2021. *models.coherencemodel – Topic coherence pipeline.* [Online] Available at: https://radimrehurek.com/gensim/models/coherencemodel.html
Řehůřek, R., 2021. *models.doc2vec - Doc2vec paragraph embeddings.* s.l.:s.n.
Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P., 2004. *The Author-Topic Model for Authors and Documents.* Arlington, Virginia, USA, AUAI Press, pp. 487-494.
Röder, M., Both, A. & Hinneburg, A., 2015. *Exploring the Space of Topic Coherence Measures.* Shanghai, China, Association for Computing Machinery.
Suominen, A. & Toivanen, H., 2016. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), pp. 2464-2476.
Tang, J., Jin, R. & Zhang, J., 2008. *A Topic Modeling Approach and Its Integration into the Random Walk Framework for Academic Search.* s.l., s.n., pp. 1055-1060.
Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M., 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), pp. 1566-1581.
Traag, V., Waltman, L. & van Eck, N. J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 03, Volume 9, p. 5233.
Wallach, H. M., 2006. *Topic Modeling: Beyond Bag-of-Words.* New York, NY, USA, Association for Computing Machinery, pp. 977-984.
Waltman, L. & Schreiber, M., 2013. On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64(2), pp. 372-379.
Waltman, L. & van Eck, N. J., 2012. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), pp. 2378-2392.
Waltman, L. & van Eck, N. J., 2015. Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, 9(4), pp. 872-894.
Yan, E., Ding, Y., Milojević, S. & Sugimoto, C. R., 2012. Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 6(1), pp. 140-153.
Yu, D. et al., 2017. Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*, 10, 12(10), pp. 1-21.

Appendix A. Details of topics with high citation impact

Topic	Keywords	Publications (full counting)	Publications (fractionalised counting)	Top 10 index
620	permanent-magnet synchronous motors, sensorless control	57	47.3	5.33
220	computer vision, local binary patterns, image classification	104	66.1	4.69
215	denoising, filtering	116	76.4	4.13
314	unmanned aerial vehicles, remote sensing	60	42.7	4.08
104	MIMO systems	158	108.6	4.01
81	millimeter wave MIMO systems	156	99.6	3.78
38	gamification, user acceptance and adoption in social media and virtual worlds	207	151.9	3.74
480	video coding and compression	73	47.7	3.72
141	RFID technologies	143	95.2	3.61
92	fading, fading channels	174	105.1	3.51
441	histone deacetylases, sirtuins	75	53.2	3.45
300	particle swarm optimization	87	61.2	3.38
124	research, publishing, scientometrics	132	93.6	3.30
1	wireless networks	558	335.5	3.23
395	DNA methylation	75	42.4	3.22
80	drug delivery	163	110.6	3.18
146	servitization, business mod-els	114	79.7	3.15
8	gut microbiota	324	198.9	3.12
2	renewable energy, energy systems	361	274.9	3.10
217	music	117	89.7	3.02

112	Arctic maritime	144	106	2.99
10	laser scanning, point clouds, LiDAR	294	239.1	2.97
189	health behaviour	92	59.4	2.94
548	CMOS, circuits	70	59.1	2.92
357	pigs, sows	77	57.5	2.91
23	nanocellulose	234	172.5	2.87
148	permanent magnet motors	136	104.5	2.79
231	photosynthesis	115	74.2	2.79
248	power converters, control systems	91	74.1	2.74
376	microalgae	77	48.9	2.71
493	collaborative learning, self-regulation, social sustainability	61	43.9	2.71
219	stochastic filtering	100	68.7	2.70
449	obesity, weight loss, bariatric surgery	67	53.9	2.68
214	cognitive networks	119	69.5	2.66
562	mm-wave and THz antennas, waveguides	59	42	2.63
430	risk management in innovations	67	50.6	2.59
389	synthetic aperture radar, remote sensing	80	51.8	2.58
533	transceivers, receivers, converters	72	54.5	2.58
195	liver diseases	125	79.8	2.48
114	(agile) software development	153	109.3	2.43
481	finite element method, eddy current and hysteresis loss	68	54.8	2.43
176	water resources	106	69.8	2.42
468	fault detection	68	48.9	2.37
442	dairy cattle	67	40.1	2.32
117	bullying, victimization	149	91	2.29
144	automation, cyber-physical security	120	81.7	2.28
88	climate change, biodiversity, species	152	80.8	2.27

26	adsorbents, adsorption from aqueous solutions	228	143.6	2.25
425	bones: formation and regeneration	69	45.6	2.22
147	multiobjective optimization	145	96.4	2.19
378	ligaments, arthroscopy	84	73.8	2.18
496	reactive oxygen species	63	40.6	2.18
299	dementia, Alzheimers	83	47.6	2.17
184	probabilistic models	126	78	2.16
14	service and customer value creation	264	204.4	2.13
242	microRNA	106	58.9	2.12
691	quartz crystal microbalance, nanocellulose	56	41.9	2.12
15	machine learning, classification	285	192.3	2.10
400	lithium-ion batteries	59	40.7	2.08
401	nanomaterials, nanoparticles	76	42.1	2.07
429	energy and housing	70	54.5	2.05
342	creativity	70	58	2.03
448	neurogenesis, brain-derived neurotrophic factor	75	48.1	2.02
170	biodiversity, conservation, protected areas	126	72.6	2.01
349	tannins and polyphenols	86	56.8	2.00

Appendix B. Research collaboration in topics with high citation impact

In the following figures, N is the number of publications, τ_{10} is the top 10 index, θ_{Int} is the proportion of international co-authored publications and θ_{Dom} is the proportion of national co-authored publications. The colours describe the citation impact. For more keywords that describe topics, see Appendix E. For the full names of the organisations, see Appendix F.

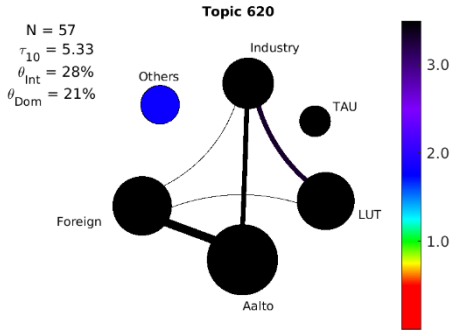


Figure B. 1. permanent-magnet synchronous motors, sensorless control

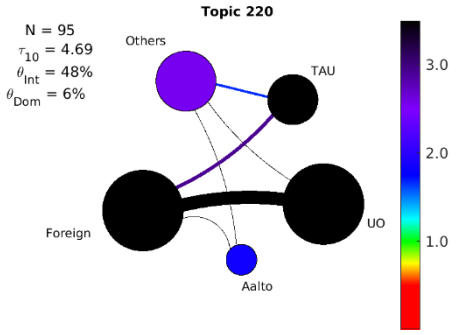


Figure B. 2. computer vision, local binary patterns, image classification

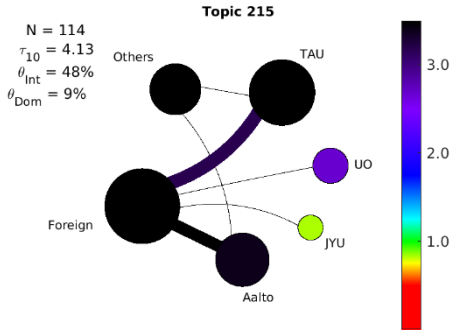


Figure B. 3. denoising, filtering

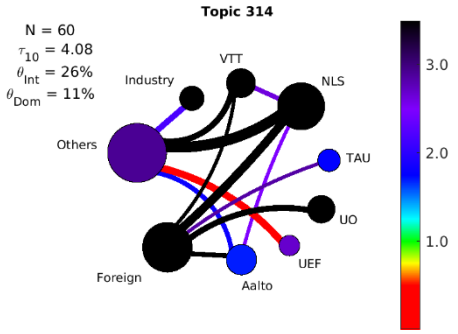


Figure B. 4. unmanned aerial vehicles, remote sensing

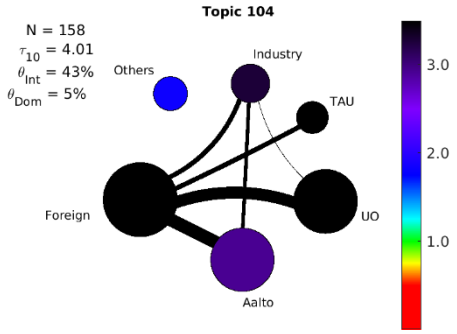


Figure B. 5. MIMO systems

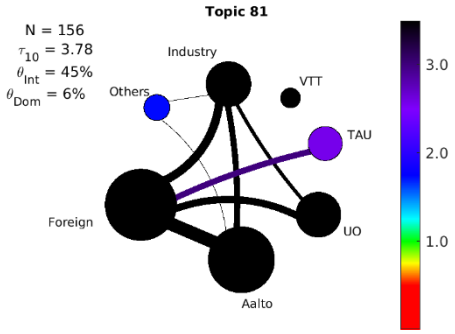


Figure B. 6. millimeter wave MIMO systems

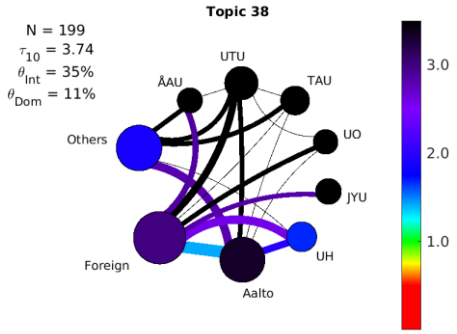


Figure B. 7. gamification, user acceptance and adoption in social media and virtual worlds

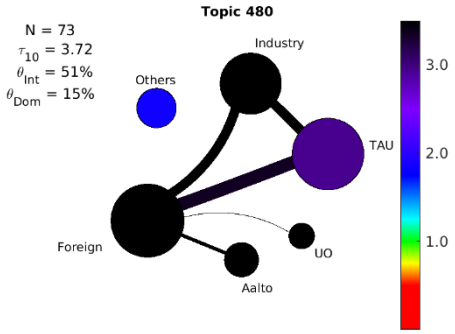


Figure B. 8. video coding and compression

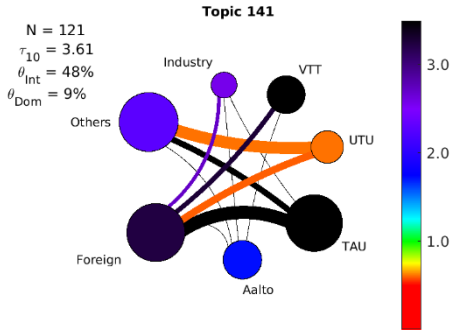


Figure B. 9. RFID technologies

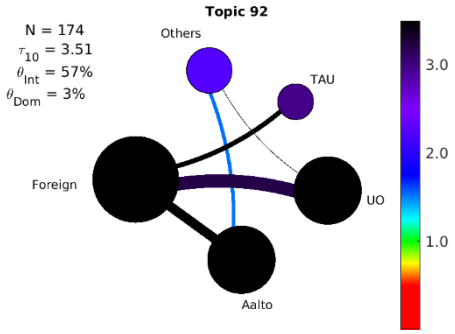


Figure B. 10. fading, fading channels

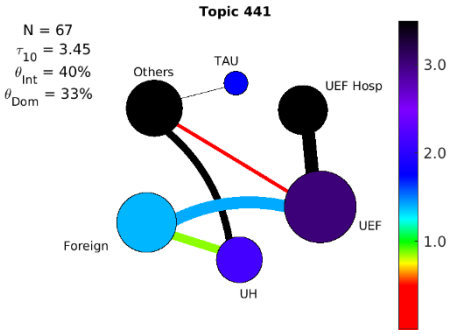


Figure B. 11. histone deacetylases, sirtuins

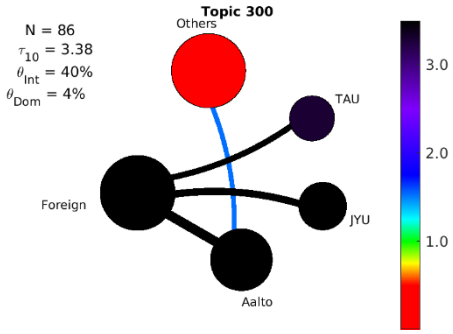


Figure B. 12. particle swarm optimization

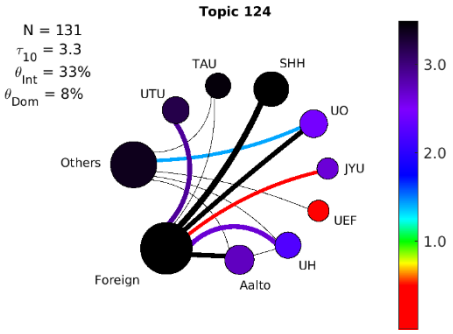


Figure B. 13. research, publishing, scientometrics

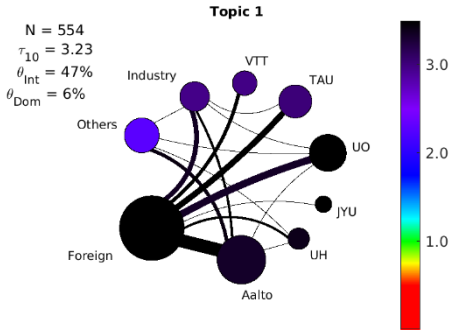


Figure B. 14. wireless networks

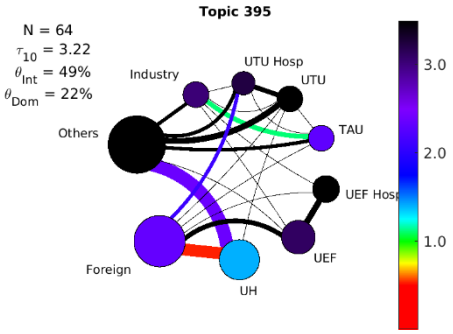


Figure B. 15. DNA methylation

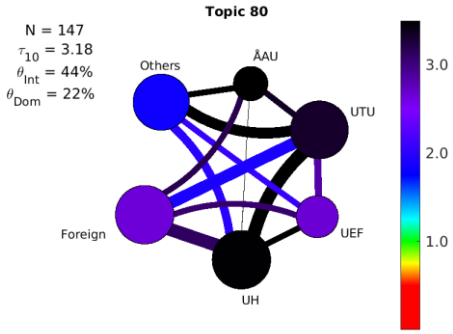


Figure B. 16. drug delivery

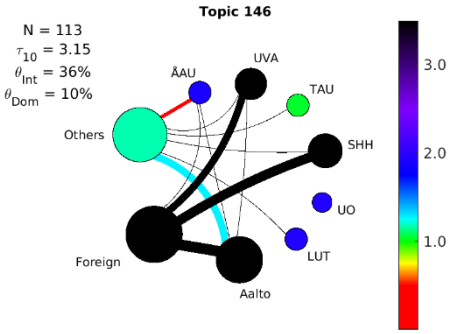


Figure B. 17. servitisation, business models

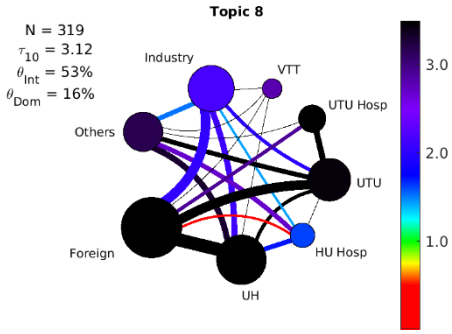


Figure B. 18. gut microbiota

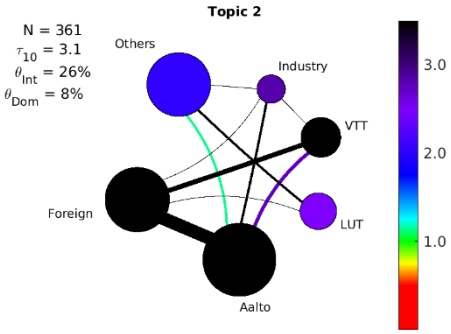


Figure B. 19. renewable energy, energy systems

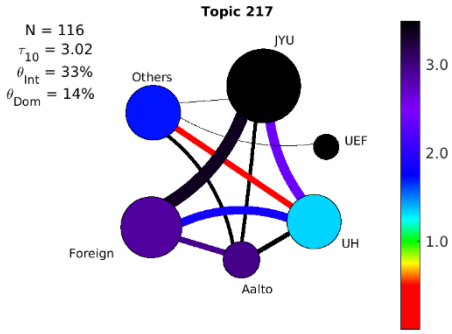


Figure B. 20. music

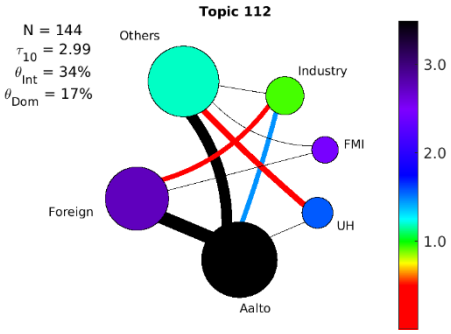


Figure B. 21. arctic maritime

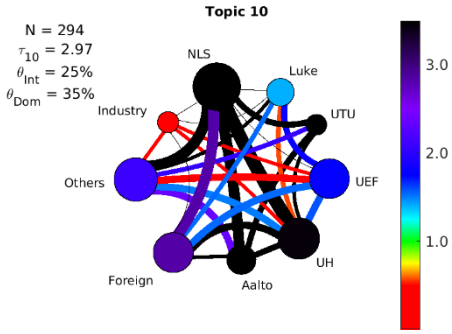


Figure B. 22. laser scanning, point clouds LiDAR

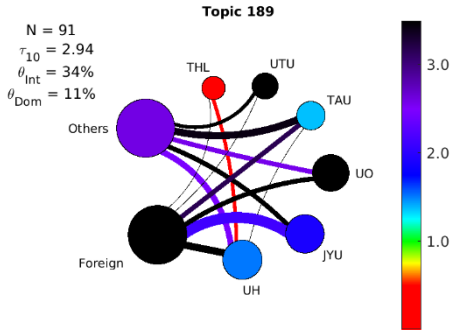


Figure B. 23. health behaviour

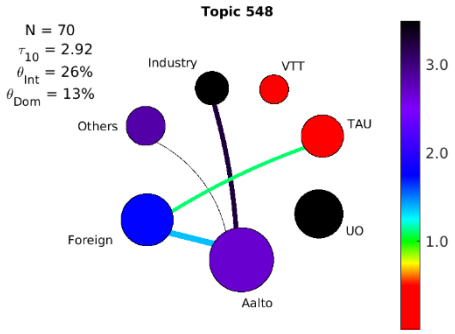


Figure B. 24. CMOS, circuits

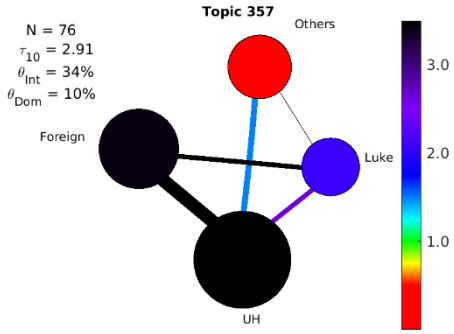


Figure B. 25. pigs, sows

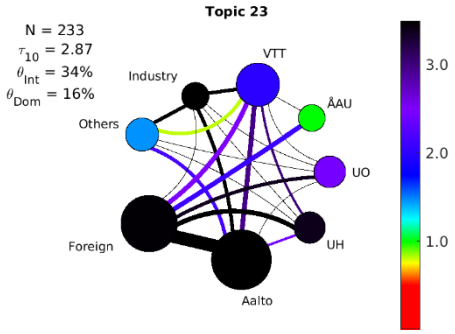


Figure B. 26. nanocellulose

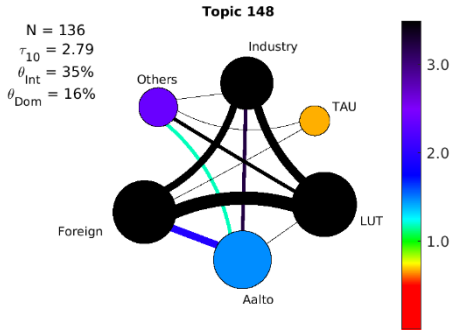


Figure B. 27. permanent magnet motors

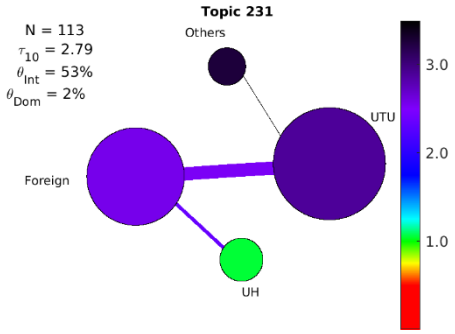


Figure B. 28. photosynthesis

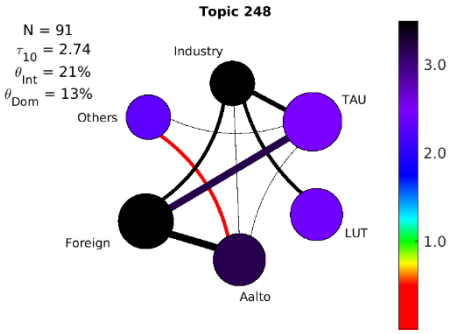


Figure B. 29. power converters, control systems

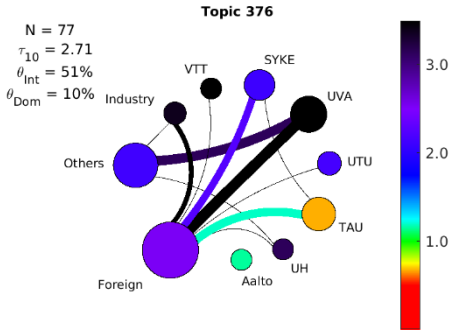


Figure B. 30. microalgae

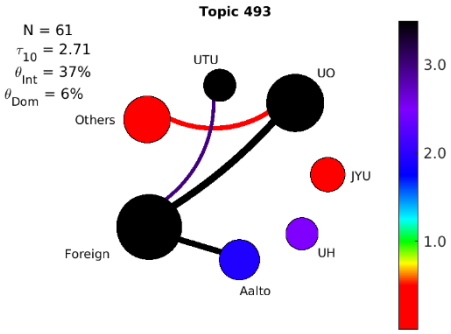


Figure B. 31. collaborative learning, self-regulation, social sustainability

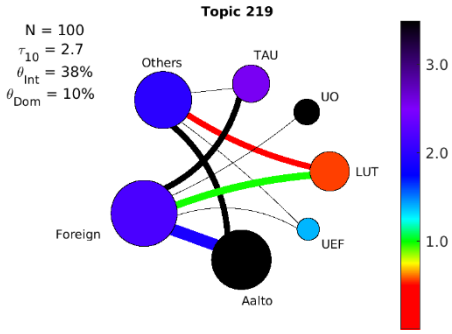


Figure B. 32. stochastic filtering

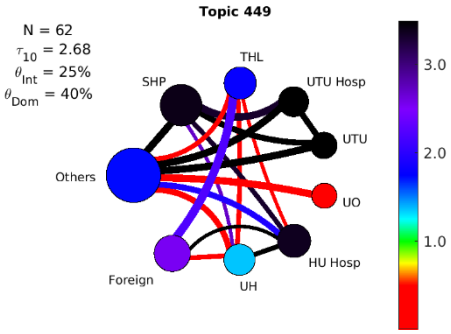


Figure B. 33. obesity, weight loss, bariatric surgery

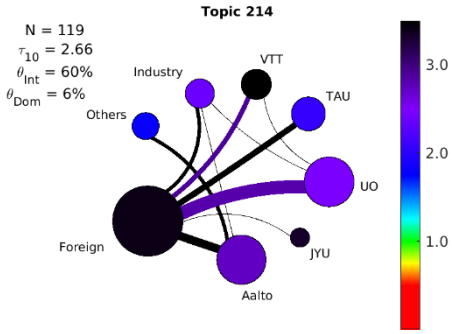


Figure B. 34. cognitive networks

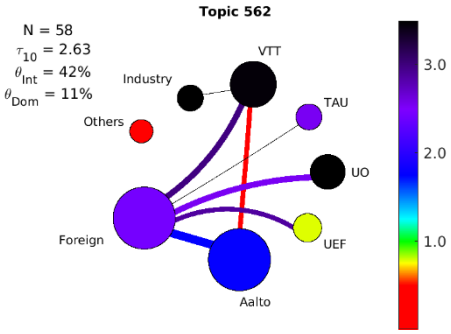


Figure B. 35. mm-wave and THz antennas waveguides

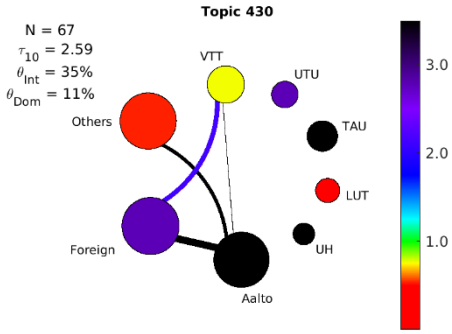


Figure B. 36. risk management in innovations

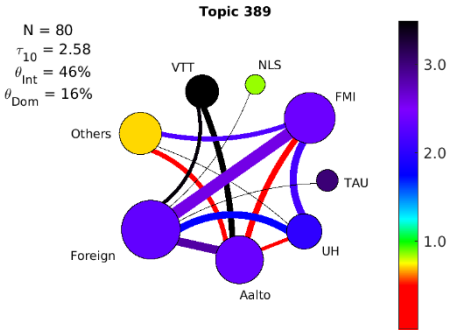


Figure B. 37. synthetic aperture radar, remote sensing

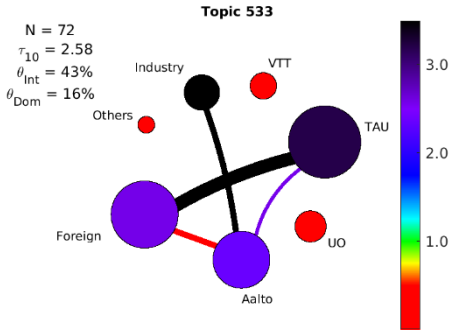


Figure B. 38. transceivers, receivers, converters

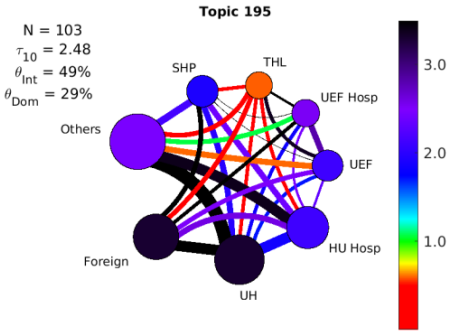


Figure B. 39. liver diseases

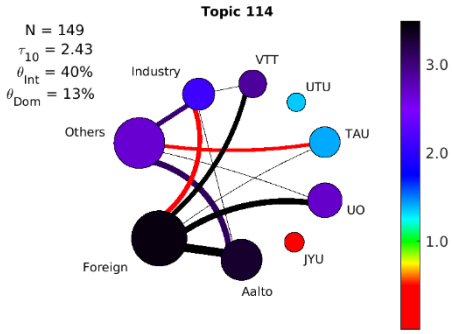


Figure B. 40. (agile) software development

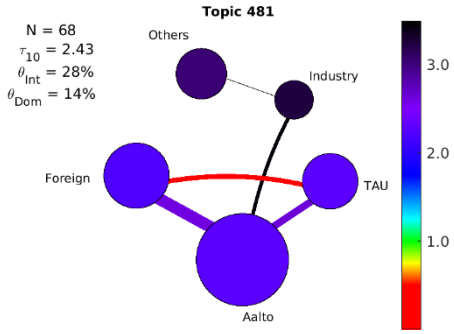


Figure B. 41. finite element method, eddy current and hysteresis loss

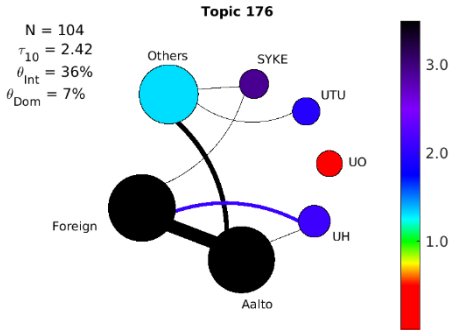


Figure B. 42. water resources

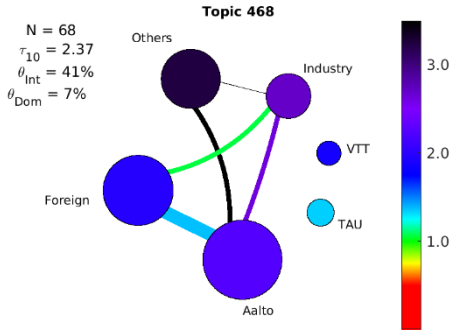


Figure B. 43. fault detection

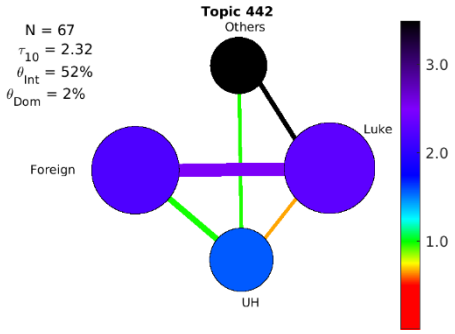


Figure B. 44. dairy cattle

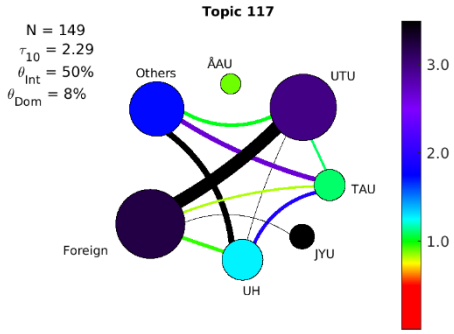


Figure B. 45. bullying, victimization

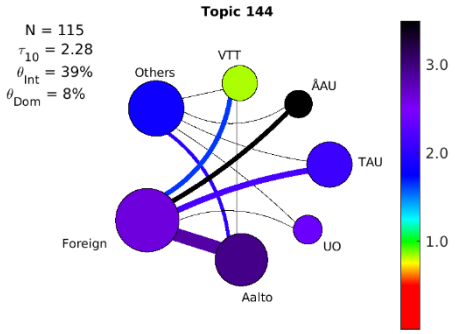


Figure B. 46. automation, cyber-physical security

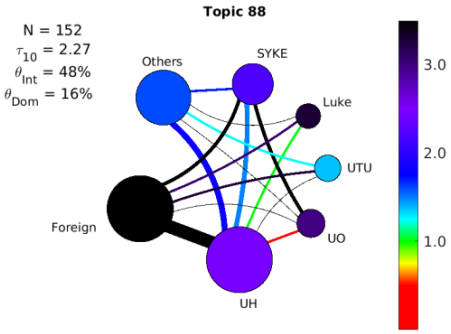


Figure B. 47. climate change, biodiversity

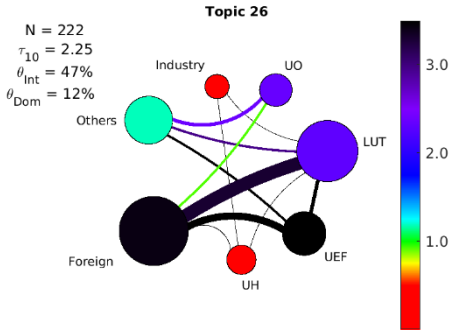


Figure B. 48. adsorbents, adsorption from aqueous solutions

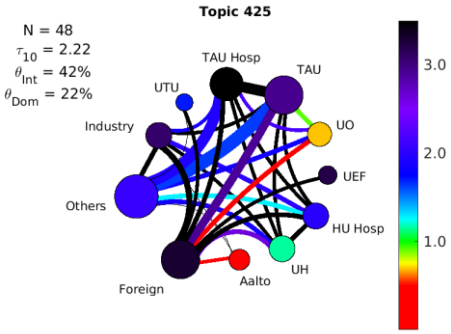


Figure B. 49. bones: formation and regeneration

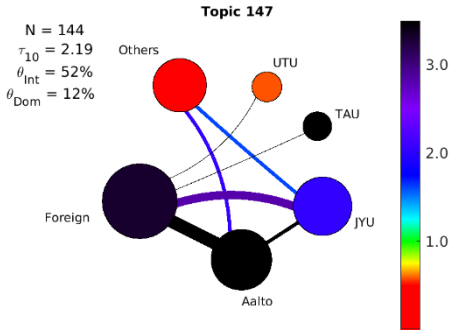


Figure B. 50. multiobjective optimization

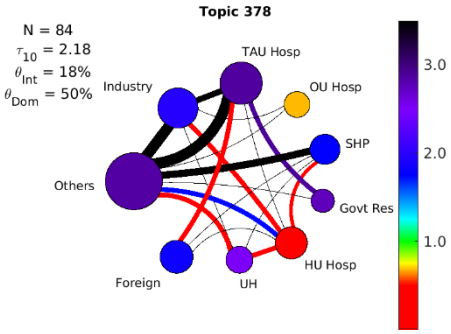


Figure B. 51. ligaments, arthroscopy

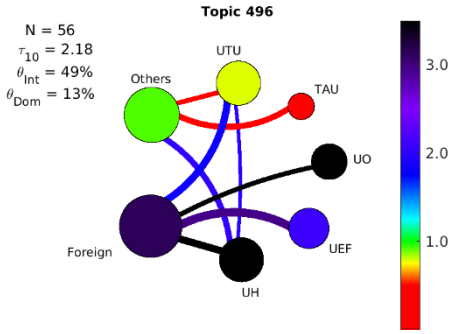


Figure B. 52. reactive oxygen species

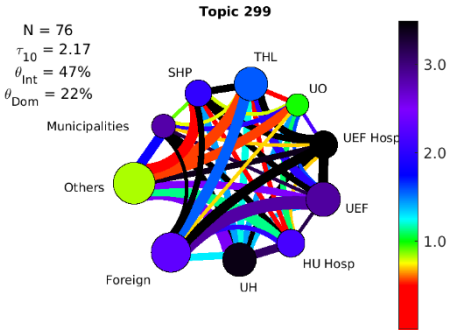


Figure B. 53. dementia, Alzheimers

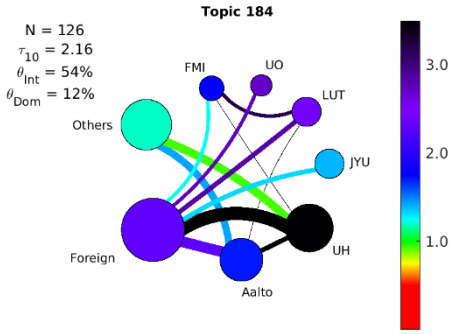


Figure B. 54. probabilistic models

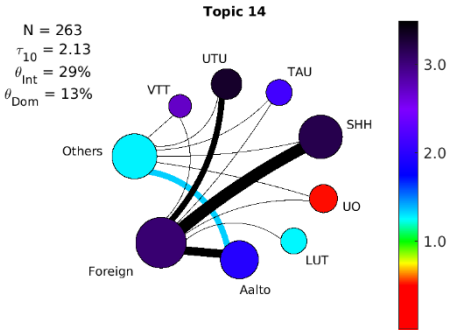


Figure B. 55. service and customer value creation

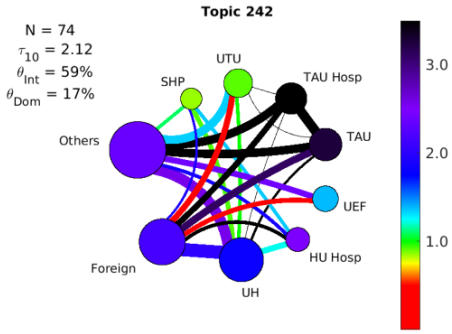


Figure B. 56. microRNA

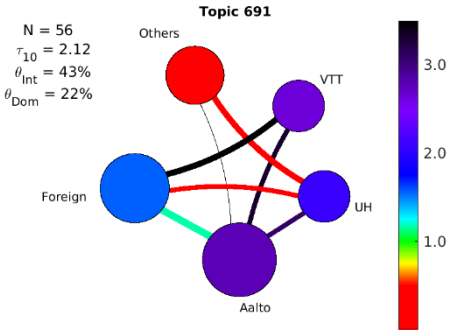


Figure B. 57. quartz crystal microbalance nanocellulose

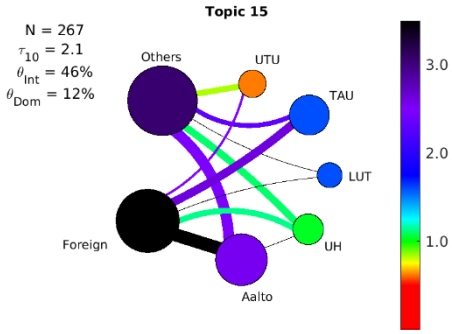


Figure B. 58. machine learning, classification

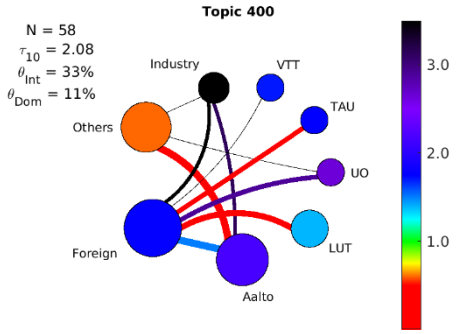


Figure B. 59. lithium-ion batteries

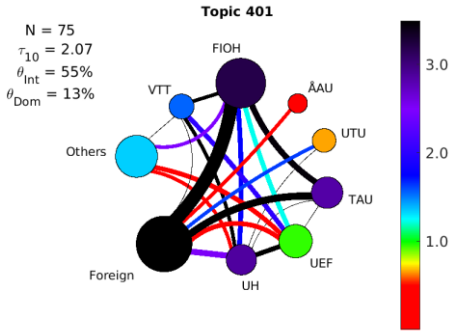


Figure B. 60. nanomaterials, nanoparticles

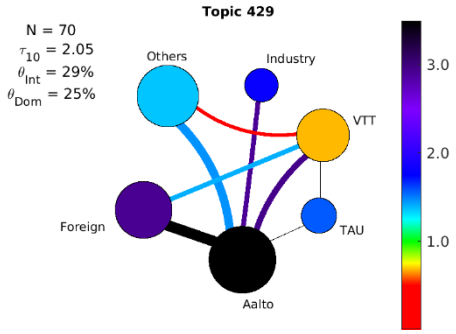


Figure B. 61. energy and housing

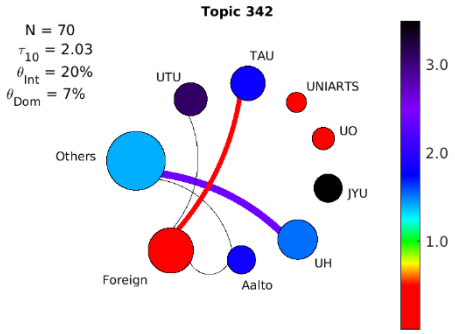


Figure B. 62. creativity

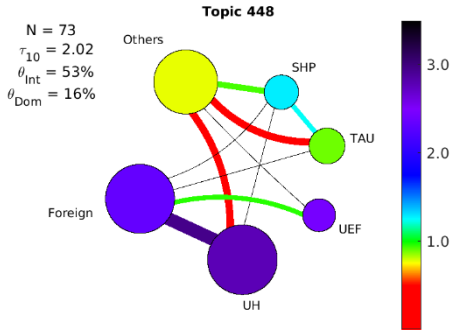


Figure B. 63. neurogenesis, brain-derived neurotrophic factor

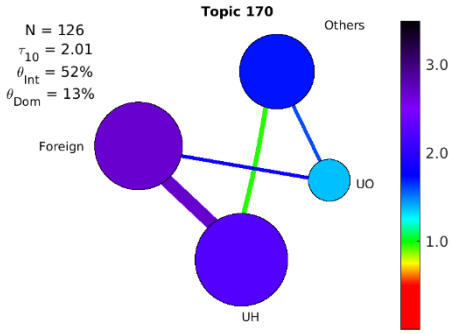


Figure B. 64. biodiversity, conservation

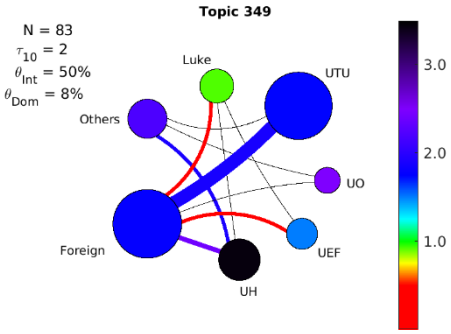


Figure B. 65. tannins and polyphenols

Appendix C. WoS subjects in topics with high citation impact

Topic	Keywords	WoS subjects	[%]
620	permanent-magnet synchronous motors, sensorless control	Engineering, Electrical & Electronic	43.3
		Engineering, Multidisciplinary	15.8
		Automation & Control Systems	12.0
		Instruments & Instrumentation	9.0
220	computer vision, local binary patterns, image classification	Computer Science, Artificial Intelligence	40.6
		Engineering, Electrical & Electronic	24.7
215	denoising, filtering	Engineering, Electrical & Electronic	41.8
		Computer Science, Artificial Intelligence	17.0
		Computer Science, Software Engineering	5.3
314	unmanned aerial vehicles, remote sensing	Remote Sensing	29.1
		Engineering, Electrical & Electronic	12.3
		Imaging Science & Photographic Technolog	10.1
		Telecommunications	5.9
		Geosciences, Multidisciplinary	5.6
104	MIMO systems	Engineering, Electrical & Electronic	50.1
		Telecommunications	37.4
81	millimeter wave MIMO systems	Engineering, Electrical & Electronic	44.6
		Telecommunications	37.8
38	gamification, user acceptance, adoption in social media & virtual world	Information Science and Library Science	21.4
		Computer Science, Information Systems	8.8
		Psychology, Multidisciplinary	8.7
		Psychology, Experimental	7.8
		Communication	7.5
		Business	6.6
480	video coding and compression	Engineering, Electrical & Electronic	52.3
		Computer Science, Information Systems	12.8
		Telecommunications	11.1
		Computer Science, Software Engineering	10.5
		Computer Science, Theory & Methods	5.9
141	RFID technologies	Engineering, Electrical & Electronic	41.9
		Telecommunications	23.6
		Instruments & Instrumentation	6.1
92	fading, fading channels	Telecommunications	49.5
		Engineering, Electrical & Electronic	37.1

Topic	Keywords	WoS subjects	[%]
441	histone deacetylases, sirtuins	Cell Biology	18.9
		Biochemistry & Molecular Biology	15.7
		Pharmacology & Pharmacy	11.5
		Geriatrics and Gerontology	9.2
		Genetics & Heredity	7.5
		Chemistry, Medicinal	6.1
300	particle swarm optimization	Computer Science, Artificial Intelligence	25.5
		Computer Science, Interdisciplinary Applica	6.9
		Computer Science, Information Systems	6.8
		Operations Research & Management Scien	6.4
		Engineering, Electrical & Electronic	5.1
124	research, publishing, scientometrics	Information Science and Library Science	16.3
		Management	11.6
		Computer Science, Interdisciplinary Applica	10.3
		Business	5.7
		Multidisciplinary Sciences	5.3
1	wireless networks	Telecommunications	28.1
		Engineering, Electrical & Electronic	19.0
		Computer Science, Information Systems	18.2
		Computer Science, Theory & Methods	5.2
395	dna methylation	Genetics & Heredity	17.3
		Geriatrics and Gerontology	9.9
		Oncology	8.5
		Biochemistry & Molecular Biology	8.2
		Immunology	6.1
		Cell Biology	6.1
		Cardiac & Cardiovascular System	5.3
80	drug delivery	Pharmacology & Pharmacy	30.8
		Materials Science, Biomaterials	8.4
		Engineering, Biomedical	8.3
		Nanoscience & Nanotechnology	8.2
		Materials Science, Multidisciplinary	7.4
		Chemistry, Multidisciplinary	6.3
146	servitization, business models	Management	34.3
		Business	28.7

Topic	Keywords	WoS subjects	[%]
8	gut microbiata	Microscopy	19.9
		Nutrition & Dietetics	19.0
		Food Science & Technology	9.2
		Gastroenterology & Hepatology	9.0
2	renewable energy, energy systems	Energy & Fuels	29.7
		Engineering, Electrical & Electronic	15.9
		Green & Sustainable Science & Technology	8.8
		Thermodynamics	6.7
		Engineering, Chemical	5.3
217	music	Psychology, Experimental	21.3
		Music	13.5
		Psychology, Multidisciplinary	6.7
112	arctic maritime	Engineering, Marine	18.0
		Engineering, Civil	12.8
		Environmental Sciences	9.1
		Operations Research & Management Scien	7.2
		Oceanography	7.1
		Engineering, Industrial	6.7
		Engineering, Environmental	6.1
10	laser scanning, point clouds, LiDAR	Forestry	34.3
		Remote Sensing	29.2
		Imaging Science & Photographic Technolog	7.2
189	health behaviour	Public, Environmental & Occupational Heal	13.8
		Computer Science, Information Systems	8.6
		Psychology, Multidisciplinary	6.6
		Medical Informatics	6.5
		Health Care Sciences & Services	6.2
548	CMOS, circuits	Engineering, Electrical & Electronic	59.5
		Instruments & Instrumentation	8.3
		Computer Science, Hardware & Architectur	7.8
		Telecommunications	6.1
357	pigs, sows	Veterinary Sciences	36.6
		Agriculture, Dairy & Animal Science	29.4
		Behavioral Sciences	6.1
		Zoology	5.5

Topic	Keywords	WoS subjects	[%]
23	nanocellulose	Materials Science, Paper & Wood	28.9
		Polymer Science	19.3
		Materials Science, Textiles	8.6
		Engineering, Chemical	6.1
148	permanent magnet motors	Engineering, Electrical & Electronic	46.3
		Physics, Applied	18.1
		Automation & Control Systems	6.5
		Instruments & Instrumentation	6.0
231	photosynthesis	Plant Sciences	47.8
		Biochemistry & Molecular Biology	15.1
		Biophysics	9.5
		Cell Biology	5.2
248	power converters, control systems	Engineering, Electrical & Electronic	66.5
376	microalgae	Biotechnology & Applied Microbiology	27.6
		Energy & Fuels	17.1
		Environmental Sciences	9.2
		Marine & Freshwater Biology	8.6
		Agricultural Engineering	7.3
493	collaborative learning, self-regulation, social sustainability	Education & Educational Research	37.6
		Psychology, Educational	12.9
		Philosophy	6.8
219	stochastic filtering	Engineering, Electrical & Electronic	31.0
		Automation & Control Systems	11.5
		Mathematics, Applied	8.0
		Computer Science, Artificial Intelligence	5.0
449	obesity, weight loss, bariatric surgery	Surgery	57.7
		Medicine, General & Internal	6.5
		Endocrinology & Metabolism	6.5
214	cognitive networks	Engineering, Electrical & Electronic	42.1
		Telecommunications	36.8
		Computer Science, Information Systems	5.9
562	mm-wave and THz antennas, waveguides	Engineering, Electrical & Electronic	44.8
		Optics	18.2
		Telecommunications	18.1
		Physics, Applied	8.5

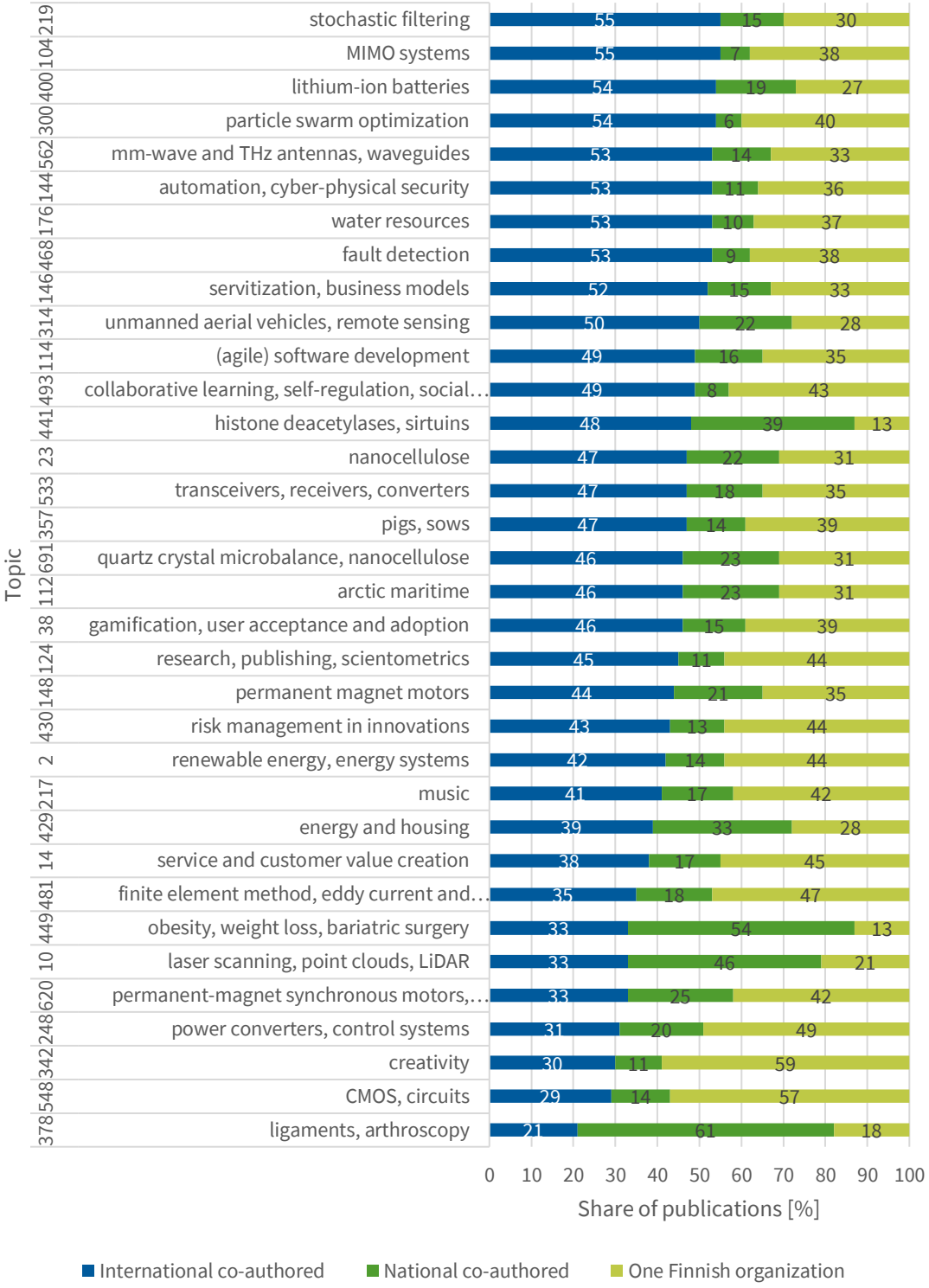
Topic	Keywords	WoS subjects	[%]
430	risk management in innovations	Management	46.9
		Business	11.2
		Social Sciences, Interdisciplinary	5.3
389	synthetic aperture radar, remote sensing	Remote Sensing	25.1
		Engineering, Electrical & Electronic	22.6
		Imaging Science & Photographic Technolog	14.4
		Geochemistry & Geophysics	9.6
		Meteorology & Atmospheric Sciences	6.6
533	transceivers, receivers, converters	Engineering, Electrical & Electronic	82.7
		Telecommunications	9.5
195	liver diseases	Gastroenterology & Hepatology	28.2
		Endocrinology & Metabolism	13.7
		Biochemistry & Molecular Biology	7.2
		Medicine, General & Internal	7.0
		Nutrition & Dietetics	5.9
114	(agile) software development	Computer Science, Software Engineering	42.0
		Computer Science, Information Systems	14.8
		Information Science and Library Science	7.2
		Computer Science, Theory & Methods	5.9
481	finite element method, eddy current and hysteresis loss	Engineering, Electrical & Electronic	39.9
		Physics, Applied	29.2
		Physics, Condensed Matter	7.8
176	water resources	Water Resources	26.8
		Environmental Sciences	16.5
		Environmental Studies	8.7
		Geography	6.0
		Geosciences, Multidisciplinary	5.9
		Engineering, Civil	5.5
468	fault detection	Engineering, Electrical & Electronic	36.2
		Automation & Control Systems	11.5
		Engineering, Mechanical	9.2
		Computer Science, Artificial Intelligence	7.7
		Engineering, Chemical	5.1

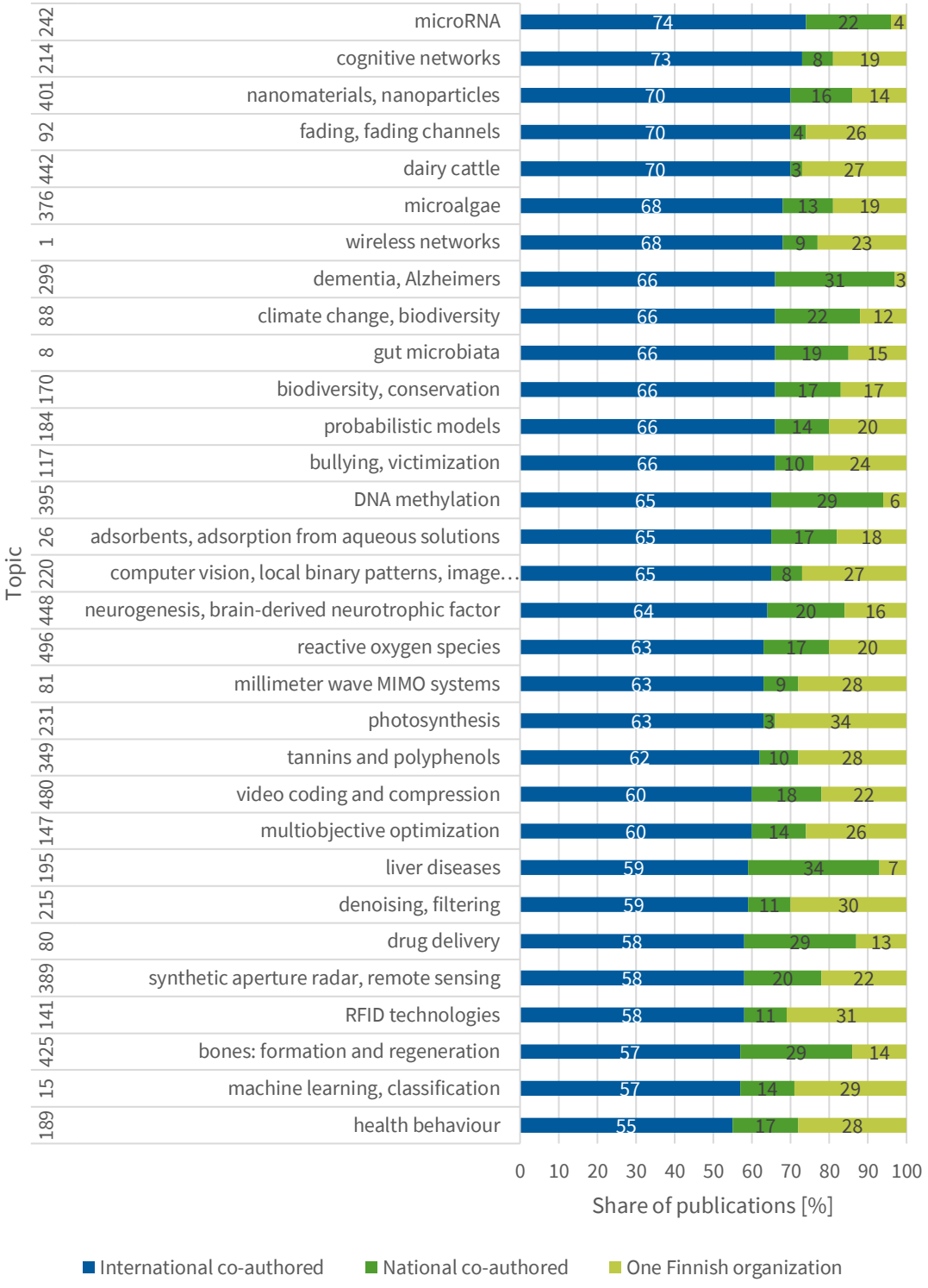
Topic	Keywords	WoS subjects	[%]
442	dairy cattle	Agriculture, Dairy & Animal Science	68.5
		Food Science & Technology	15.0
		Veterinary Sciences	8.7
117	bullying, victimization	Psychology, Developmental	21.0
		Education & Educational Research	15.9
		Psychology, Educational	8.0
		Psychology, Social	6.6
		Criminology & Penology	6.2
		Psychiatry	6.0
		Psychology, Multidisciplinary	5.8
144	automation, cyber-physical security	Computer Science, Information Systems	11.4
		Engineering, Electrical & Electronic	10.6
		Computer Science, Software Engineering	8.9
		Computer Science, Artificial Intelligence	8.9
		Telecommunications	6.2
		Automation & Control Systems	5.5
88	climate change, biodiversity	Ecology	31.3
		Biodiversity Conservation	18.5
		Environmental Sciences	14.0
		Geography, Physical	5.9
26	adsorbents, adsorption from aqueous solutions	Engineering, Chemical	18.7
		Environmental Sciences	15.4
		Engineering, Environmental	11.9
		Chemistry, Physical	8.2
		Chemistry, Inorganic & Nuclear	6.1
		Water Resources	6.0
425	bones: formation and regeneration	Materials Science, Biomaterials	12.2
		Multidisciplinary Sciences	11.7
		Cell and Tissue Engineering	10.9
		Engineering, Biomedical	10.4
		Cell Biology	8.1
147	multiobjective optimization	Operations Research & Management Scien	22.4
		Computer Science, Artificial Intelligence	9.5
		Mathematics, Applied	8.0
		Computer Science, Interdisciplinary Applic	6.9
		Engineering, Multidisciplinary	5.1

Topic	Keywords	WoS subjects	[%]
378	ligaments, arthroscopy	Orthopedics	32.2
		Sport Sciences	25.3
		Surgery	14.7
		Radiology, Nuclear Medicine & Medical Ima	10.2
496	reactive oxygen species	Biochemistry & Molecular Biology	17.3
		Plant Sciences	17.2
		Cell Biology	10.1
		Toxicology	7.6
		Cardiac & Cardiovascular System	5.1
299	dementia, Alzheimers	Geriatrics and Gerontology	27.5
		Neurosciences	22.6
		Clinical Neurology	13.6
		Gerontology	6.0
		Medicine, General & Internal	5.8
184	probabilistic models	Statistics & Probability	20.8
		Computer Science, Artificial Intelligence	7.2
		Ecology	5.3
14	service and customer value creation	Business	39.3
		Management	36.5
242	microRNA	Oncology	19.7
		Biochemistry & Molecular Biology	11.0
		Cell Biology	10.7
		Genetics & Heredity	7.6
		Pathology	5.5
691	quartz crystal microbalance, nanocellulose	Chemistry, Physical	25.6
		Materials Science, Paper & Wood	17.8
		Polymer Science	7.7
		Chemistry, Analytical	5.6
15	machine learning, classification	Computer Science, Artificial Intelligence	41.7
		Engineering, Electrical & Electronic	10.5
		Computer Science, Information Systems	6.1

Topic	Keywords	WoS subjects	[%]
400	lithium-ion batteries	Energy & Fuels	17.4
		Chemistry, Physical	14.0
		Materials Science, Multidisciplinary	13.5
		Metallurgy & Metallurgical Engineering	8.6
		Engineering, Electrical & Electronic	7.2
		Electrochemistry	6.6
401	nanomaterials, nanoparticles	Toxicology	41.1
		Nanoscience & Nanotechnology	10.4
		Environmental Sciences	7.1
429	energy and housing	Energy & Fuels	26.8
		Construction & Building Technology	24.6
		Engineering, Civil	15.4
		Public, Environmental & Occupational Heal	7.3
		Green & Sustainable Science & Technology	5.7
342	creativity	Education & Educational Research	23.9
		Management	12.3
		Art	8.6
		Information Science and Library Science	8.4
448	neurogenesis, brain-derived neurotrophic factor	Neurosciences	48.2
		Pharmacology & Pharmacy	10.4
		Clinical Neurology	7.2
		Psychiatry	7.1
170	biodiversity, conservation	Ecology	26.8
		Biodiversity Conservation	22.6
		Environmental Sciences	17.9
349	tannins and polyphenols	Food Science & Technology	14.7
		Plant Sciences	12.4
		Biochemistry & Molecular Biology	10.3
		Chemistry, Applied	9.5
		Ecology	8.8
		Agriculture, Multidisciplinary	7.0

Appendix D. National and international collaboration in topics with high citation impact




Topic
242 microRNA 74 22 4
214 cognitive networks 73 8 19
401 nanomaterials, nanoparticles 70 16 14
92 fading, fading channels 70 4 26
442 dairy cattle 70 3 27
376 microalgae 68 13 19
1 wireless networks 68 9 23
299 dementia, Alzheimers 66 31 3
88 climate change, biodiversity 66 22 12
8 gut microbiata 66 19 15
170 biodiversity, conservation 66 17 17
184 probabilistic models 66 14 20
117 bullying, victimization 66 10 24
395 DNA methylation 65 29 6
26 adsorbents, adsorption from aqueous solutions 65 17 18
220 computer vision, local binary patterns, image… 65 8 27
448 neurogenesis, brain-derived neurotrophic factor 64 20 16
496 reactive oxygen species 63 17 20
81 millimeter wave MIMO systems 63 9 28
231 photosynthesis 63 3 34
349 tannins and polyphenols 62 10 28
480 video coding and compression 60 18 22
147 multiobjective optimization 60 14 26
195 liver diseases 59 34 7
215 denoising, filtering 59 11 30
80 drug delivery 58 29 13
389 synthetic aperture radar, remote sensing 58 20 22
141 RFID technologies 58 11 31
425 bones: formation and regeneration 57 29 14
15 machine learning, classification 57 14 29
189 health behaviour 55 17 28
0 10 20 30 40 50 60 70 80 90 100
Share of publications [%]
International co-authored
National co-authored
One Finnish organization

Appendix E. Keywords of topics with high citation impact

The following tables show a summary of the keywords in topics with high citation impact. The column “Model keywords” refers to the 25 most significant output keywords of the topic model. To clarify the key ideas of the topics, topics are labelled based on the model keywords by an analyst without subject matter expertise. These label keywords are shown in the column “Human-labelled keywords” and are used throughout the report. In addition, the term frequency–inverse document frequency (tf-idf) weights were calculated to find the most important words for each topic based on the titles, abstracts and keywords of the publications. The most important words found are presented in the column “tf-idf keywords”.

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
620	permanent-magnet synchronous motors, sensorless control	sensorless drives synchronous motors stator machines reluctance induction_motor magnet rotor torque permanent observer inverter winding induction drive speed machine harmonics inductance motor converters controller paper_deals	control speed motor drives induction variable
220	computer vision, local binary patterns, image classification	local_binary binary_patterns descriptors feature texture classification binary support_vector pattern face image lbp recognition vector_machine discriminant convolutional deep_learning representation automatic images vector_machines features automatically local unsupervised	texture local recognition face image binary
215	denoising, filtering	denoising sparse sparsity noisy filtering regularization nonlocal iterative noise convolutional estimation algorithm image gaussian subspace algorithms kernel filters signals wavelet filter squares robust feature transforms	noise image denoising sparse based signal
314	unmanned aerial vehicles, remote sensing	uav unmanned_aerial unmanned aerial aerial_vehicle photogrammetry hyperspectral point_clouds radiometric point_cloud remote_sensing laser_scanner imagery camera from remote sensing_data lidar cameras scanner laser_scanning airborne multispectral vehicles sensing	aerial uav unmanned hyperspectral unmanned_aerial based

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
104	MIMO systems	mimo multiuser downlink beamforming input_multiple transmit mimo_systems output_mimo maximization input massive_mimo output transceiver ofdma multiple fading fading_channels multiple_access ofdm wireless relaying interference relay communications orthogonal	multiple mimo multi power input networks
81	millimeter wave MIMO systems	millimeter millimeter_wave ghz mimo antenna multipath output_mimo communications mmwave input_multiple wireless wideband transceiver backhaul beamforming antennas mimo_systems input radio propagation output channel massive_mimo fading_channels fading	wave multiple antenna millimeter input channel
38	gamification, user acceptance and adoption in social media and virtual worlds	user_acceptance gratifications continuance virtual_worlds information acceptance commerce adoption usage social_media worlds gamification online intention user facebook ease networking technology purchase intentions internet virtual consumer continued	technology social mobile information self adoption
480	video coding and compression	video_coding video coding streaming scalable multimedia bit mobile_devices quantization dvb architectures proposes audio edge_computing resilient computing wireless scheme aware broadcast paper_proposes algorithms receivers sparse cloud_computing	video coding video_coding streaming rate depth
141	RFID technologies	rfid rfid_tag radio_frequency tag uhf chipless tags antennas wearable antenna textile passive printable sensor frequency rf inkjet inks radio electronics mhz printed conductive wireless sensors	rfid frequency tag antenna radio sensor
92	fading, fading channels	fading outage fading_channels relaying relay decode amplify nakagami mimo input_multiple transmit output_mimo forward channels channel interference wireless duplex mimo_systems communications probability cooperative secrecy multiuser orthogonal	duplex relay forward fading channel multiple
441	histone deacetylases, sirtuins	sirtuins sirt deacetylase histone caenorhabditis elegans nad lysine span ampk longevity nf senescence homeostasis transcription lifespan translational acetylation aging kappa ribose chromatin adp chaperone parp	elegans kappa caenorhabditis span life protein

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
300	particle swarm optimization	swarm particle_swarm harmony heuristic optimization algorithm optimisation multiobjective algorithms pareto heuristics nonsmooth paper_proposes inspired robust programming computationally systems_paper neural_networks multi proposes interactive noisy clustering modal	optimization algorithm search based harmony swarm
124	research, publishing, scientometrics	journals publishing open_access articles scholarly citation journal publication scientific publications published reviewed research funding science sciences literature social_sciences oa charges authors academia ranking author researchers	research open access journals open_access science
1	wireless networks	wireless sensor_networks edge_computing internet_iot ad_hoc hoc_networks iot wireless_sensor virtualization networks communications vehicular cloud_computing hoc authentication computing mobile_networks privacy secure sensor_network fog network qos aware ieee	networks mobile network wireless internet based
395	DNA methylation	dna_methylation methylation epigenetic epigenetics epigenome cpg noncoding histone promoter chromatin suppressor gene_expression genes cancer_dna dna genome gene transcription transcriptional wide rnas gwas microrna transcriptome aberrant	methylation dna dna_methylation expression epigenetic gene
80	drug delivery	drug_delivery porous_silicon delivery nanocarriers mesoporous cellular_uptake biodistribution nanoparticles cancer_therapy drug porous microparticles drug_release liposomes gene_delivery silicon liposome nanoparticle responsive vivo plga doxorubicin cytotoxicity targeting loaded	delivery drug silicon nanoparticles porous porous_silicon
146	servitization, business models	servitization offerings service_systems product business_models capabilities dominant_logic creation service customer business value_crea-	service business product servitization innovation manufacturing

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
		tion business_model value_co dominant innovation propositions methodology advantage approach_paper customization digitalization based_view marketing customer_value	
8	gut microbiota	microbiota gut gut_microbiota probiotics bifidobacteria human_gut probiotic prebiotics lactobacillus bifidobacterium rhamnosus acidophilus akkermansia intestinal muciniphila irritable gg microbiome rhamnosus_gg gut_microbiome microflora butyrate mucus lactobacilli prebiotic	microbiota lactobacillus gut human intestinal probiotics
2	renewable energy, energy systems	energy_system electricity energy_storage wind_power side_management energy_systems demand_response renewable power_systems combined_heat power_system solar_power energy_sources heat_power chp demand grids heat_pump power smart_grid grid microgrids storage heating cogeneration	energy power renewable demand heat systems
217	music	music musical emotion listening musicians perception emotions emotional sad affective auditory brain_responses dance naturalistic aesthetics fmri sounds auditory_cortex human_auditory neuroscience perceptual mood stimuli potential_erp stimulus	music musical emotion emotions performance perception
112	arctic maritime	ship maritime ships spill shipping risk_analysis oil_spill ais gulf_finland gulf accidents collision passenger ice transportation navigation probabilistic traffic accident autonomous sea loads flooding port belief	ship ice risk oil maritime sea
10	laser scanning, point clouds, LiDAR	laser_scanning airborne_laser scanning_als lidar individual_tree laser_scanner tree_detection scanning_data stem_volume point_clouds airborne airborne_lidar lidar_data scanning point_cloud mobile_laser scanner tree_height attributes photogrammetry als plot tls neighbor remote_sensing	forest laser scanning laser_scanning tree based

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
189	health behaviour	behavior_change health_behavior theory_planned planned intentions persuasive interventions habit gamification intention acceptance motivational behaviors exercise_health behaviours ehealth user_acceptance promotion motivation enjoyment support_systems continuance contextual mindfulness lifestyle	self behavior health activity physical theory
548	CMOS, circuits	cmos oscillator amplifier converter receiver amplifiers locked rf receivers phase_noise converters ghz voltage circuits switched wideband circuit noise nm band avalanche front transceiver analog digital	time digital phase cmos noise based
357	pigs, sows	sows sow pigs biting animal_welfare pig herds swine housed lameness welfare lactation slaughter insemination oxytocin tail housing prolactin cortisol animal farm herd lactating farms dairy_cows	tail welfare sows pigs animal biting
23	nanocellulose	nanofibrillated cellulose nfc nanofibrils nanocellulose cellulose_nfc mfc suspensions microfibrils nanofibrillar rheology tempo fibers rheological cellulosic aerogels flocculation nanofibers cationic polyelectrolyte viscoelasticity bleached cnf papermaking hydrogels	cellulose properties films nanocellulose water nanofibrillated
148	permanent magnet motors	machines magnet motors permanent winding synchronous stator rotor torque induction_motor machine inductance drives losses induction reluctance sensorless speed traction fem harmonics inverter generators drive eccentricity	permanent magnet machines speed high element
231	photosynthesis	photosystem thylakoid photosystem_ii chloroplasts photoinhibition photosynthetic chlamydomonas reinhardtii synechocystis sp_pcc photosynthesis thaliana chloroplast pcc_ arabidopsis ferredoxin chlorophyll pcc acclimation nadp ndh singlet phosphorylation cyanobacterium plastid	photosystem light protein electron arabidopsis thaliana

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
248	power converters, control systems	converter converters dc inverter voltage capacitor switched power circuit boost power_system grid drives power_systems cmos capacitors generator amplifier proposes controller grids sensorless ac connected photovoltaic	power control mode voltage converter grid
376	microalgae	chlorella microalgae microalgal biodiesel cultivation algae biogas biofuel wastewater algal vulgaris anaerobic biofuels digestion alga feedstock biohydrogen sludge wastewaters biomass waste biorefinery manure water_treatment nutrient	production microalgae water chlorella wastewater lipid
493	collaborative learning, self-regulation, social sustainability	socially_shared socially collaborative shared learning classroom regulated situations motivation student problem_solving regulation pedagogical scripts school_students teacher collaboration emotions students based_learning instructional achievement teaching engagement situational	learning regulation self collaborative shared collaborative_learning
219	stochastic filtering	kalman kalman_filter filter filtering estimation filters bayesian mcmc bayes rao gaussian parameter metropolis smoothing ensemble markov chain_monte markov_chain variational linear iteration approximation discretization stochastic approximations	kalman filter kalman_filter estimation gaussian bayesian
449	obesity, weight loss, bariatric surgery	bariatric gastric_bypass roux morbid gastrectomy en weight_loss obese gastric bypass obesity surgery laparoscopic laparoscopy loss weight esophagus body_weight liver_fat bypass_surgery endoscopy preoperative risk_patients postoperative term_results	surgery weight loss bariatric bariatric_surgery obesity
214	cognitive networks	cognitive_radio radio_networks spectrum_access wireless communications radio opportunistic lte networks_paper networks hoc_networks interference spectrum licensed cooperative access fading fading_channels ad_hoc allocation multiuser ofdm multipath relaying receivers	spectrum radio cognitive cognitive_radio sensing networks

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
562	mm-wave and THz antennas, waveguides	millimeter thz ghz waveguide antenna millimeter_wave mm antennas band metamaterial wideband terahertz wave rod dielectric propagation waveguides permittivity broadband array submillimeter slot absorbers wafer nm	antenna wave waveguide millimeter thz slot
430	risk management in innovations	projects innovation risk_management new_product project front decision open_innovation capabilities strategic strategy managing product portfolio innovations organizational methodology management publication based_view value_creation business managerial firm conceptual	management innovation development project risk end
389	synthetic aperture radar, remote sensing	aperture_radar sar radar_sar aperture radar polarimetric snow remote_sensing radiometer radiometry band interferometric synthetic sea_ice interferometry polarimetry ice_thickness backscattering albedo soil_moisture satellite smos remote land_cover modis	radar sar band snow ice aperture
533	transceivers, receivers, converters	receivers adc analog amplifier sigma receiver rf amplifiers cmos converters converter linearization wideband front digital quadrature modulator transceiver compensation delta filters nonlinearity filtering linearity ghz	sigma digital delta analog direct frequency
195	liver diseases	fatty_liver steatosis alcoholic_fatty liver_disease steatohepatitis nonalcoholic alcoholic nafld liver_fat pnpla liver hepatic cirrhosis triglyceride insulin nash triacylglycerol hepatocellular fatty adipose sf metabolic glutamyl glucose_insulin fat	liver disease non alcoholic fatty insulin
114	(agile) software development	agile software projects enterprise engineering new_product customization conceptual deployment teams product development lifecycle design requirements knowledge elicitation project innovation technology methodology adopting methodologies challenges information	software development agile design engineering product

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
481	finite element method, eddy current and hysteresis loss	hysteresis fem finite losses machines finite_element stepping element laminated element_methods eddy electrical currents ferromagnetic induction_motor current sheets motors rotor finite_elements winding machine stator computation vibrations	element finite magnetic eddy model electrical
176	water resources	water_resources mekong river_basin hydropower basin river nexus hydrological flood basins sap change_impacts southeast hydrology scarcity impacts vulnerability irrigation asia food_security rivers recharge catchment climate_change aquifer	water climate change river basin climate_change
468	fault detection	fault_detection fault faults fault_diagnosis mv wavelet paper_proposes fuzzy_logic inverter proposes induction_motor transformer failures voltage fuzzy motors bearings rotor neural_networks compensated power_system convolutional detection machine eccentricity	fault detection arc based distribution process
442	dairy cattle	holstein dairy_cattle cattle dairy cows dairy_cows dairy_cow lactation feed_intake cow beef mastitis bulls lactating feed lameness silage insemination grass_silage milk herds breed breeds carcass body_condition	cattle dairy genetic genomic milk dairy_cattle
117	bullying, victimization	victimization bullying victims harassment peer victim aggression adolescence peers school esteem adolescents internalizing school_students childrens youth friendship social_network antisocial psychosocial school_children delinquency behaviors school_burnout classroom	bullying school social victimization peer children
144	automation, cyber-physical security	cyber automation ontology smart semantic execution infrastructures software architecture factory internet_iot iot iec intelligent customization smart_grid computing system_design lifecycle systems platform distributed cloud_computing internet security	cyber systems based physical design knowledge
88	climate change, biodiversity, species	range_shifts global_change climate_change extinction_risk protected_areas effects_climate climate change_impacts biodiversity change species_traits habitat conservation change_climate landscape	change climate species land conservation climate_change

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
		global_warming projected richness land habitat_loss butterflies use_change ecological abundance global_climate	
26	adsorbents, adsorption from aqueous solutions	adsorbents adsorbent isotherm aqueous biosorption adsorption removal ni_ii metal_ions adsorptive methylene isotherms sorption solution water_treatment wastewaters ions chelating cu_ii ion_exchange chitosan hexavalent metals sorbent dyes	adsorption water aqueous metal removal solution
425	bones: formation and regeneration	osteogenic human_adipose adipose_stem stem_cells stromal_cells bone_tissue osteoblast scaffolds stromal mesenchymal stem bioactive_glass osteoblasts regenerative scaffold marrow bone_marrow adipose bone tissue differentiation stem_cell hydrogel cell_therapy hydroxyapatite	cells stem bone cell tissue stem_cells
147	multiobjective optimization	multiobjective pareto optimization optimality interactive algorithms multicriteria programming computationally algorithm expensive integer decision_making objective nonsmooth decision evolutionary optimal minimization convex surrogate optimisation heuristic solving multi	optimization objective multiobjective multiobjective_optimization multi interactive
378	ligaments, arthroscopy	ligament cruciate patellar patellofemoral knee hamstring anterior arthroscopic tendon medial dislocation bundle femoral ankle osteoarthritis injuries tibial fractures arthroscopy tears biomechanical biomechanics knee_joint arthroplasty total_knee	ligament bundle knee dislocation follow cruciate
496	reactive oxygen species	reactive_oxygen oxygen_species ros species_ros nadph oxidative superoxide dismutase oxygen peroxide redox glutathione salicylic nrf thioredoxin thaliana arabidopsis cell_death signalling antioxidants peroxidation radicals reactive abscisic cytosolic	oxygen reactive species reactive_oxygen ros oxygen_species

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
299	dementia, Alzheimers	impairment dementia alzheimers vascular_risk alzheimer mild_cognitive disease_ad midlife cognitive mci old_age geriatric decline older mild cognition older_adults frailty disease_amyloid apoe ad adults mortality_older brain_atrophy apolipoprotein	disease dementia risk alzheimer cognitive alzheimers
184	probabilistic models	mcmc chain_monte markov_chain monte carlo monte_carlo metropolis markov bayesian inference likelihood graphical bayes estimation gibbs unbiased parameter stochastic approximation statistical kalman serpent kalman_filter estimating computation	carlo monte markov monte_carlo bayesian chain
14	service and customer value creation	dominant_logic creation customer value_co value_creation customer_value dominant propositions perceived_value offerings service marketing servitization business service_systems business_models value logic approach_paper methodology service_design product capabilities innovation conceptual	service value creation business customer dominant
242	microRNA	mirna mirnas micrornas microrna mir rnas transcriptional transcription mrna suppressor noncoding microarray androgen differentially gene_expression regulators cancer_cells chromatin rna expression cancer_cell messenger adipocyte castration transcriptome	mir cancer expression microrna mirna gene
691	quartz crystal microbalance, nanocellulose	qcm microbalance quartz_crystal quartz polyelectrolyte multilayers cationic afm cellulose cellulose_nfc nanofibrils nfc nanofibrillar plasmon adsorption films ultrathin atomic_force dissipation anionic carboxymethyl adsorbed mfc nanofibrillated chitosan	cellulose qcm quartz crystal films adsorption
15	machine learning, classification	support_vector classifiers vector_machines classifier vector_machine supervised machine kernel clustering feature discriminant unsupervised classification dimensionality neural_networks regression neural learning vector algorithms neural_network bayes speaker robust prediction	learning data classification analysis machine based

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
400	lithium-ion batteries	batteries lithium ion_batteries li battery cathode titanate supercapacitor electrochemical binder electrode ion anode electrodes electrolytes ti graphite electrolyte hydride energy_storage graphene_oxide cobalt milling solvent redox	ion lithium battery batteries energy carbon
401	nanomaterials, nanoparticles	nanomaterials genotoxicity inhalation nanotubes nanomaterial engineered toxicity toxicological cytotoxicity walled nanoparticle walled_carbon asbestos toxicology health_effects exposure_health nanoparticles exposure cnt nanotube comet carbon_nanotube bronchial manufactured urban_air	carbon nanomaterials nanoparticles exposure nanotubes toxicity
429	energy and housing	renovation buildings heat_pump houses building residential savings heating cold_climate demand comfort thermal_comfort energy_demand indoor cogeneration electricity office district saving energy_saving renewable energy house demand_response exergy	energy buildings efficiency building cost energy_efficiency
342	creativity	creativity creative knowledge innovation conceptual perspective organizational contexts organizations open_innovation thinking collaborative learning experiential context leadership approach_data research_design information research orchestration creation pedagogical research_agenda sensemaking	creativity innovation design creative learning work
448	neurogenesis, brain-derived neurotrophic factor	bdnf trkb neurotrophic fluoxetine neurogenesis antidepressant dentate_gyrus hippocampus mood_disorders synaptic gyrus dentate val prefrontal neurons potentiation hippocampal met knockout_mice neuroprotection rat_hippocampus ampa brain reuptake forebrain	brain bdnf derived factor depression plasticity
170	biodiversity, conservation, protected areas	prioritization protected_areas conservation zonation protected biodiversity hotspots reserve landscapes planning priority landscape areas land complementarity threatened habitat priorities ecosystem habitats habitat_loss ecological area_network richness reserves	conservation biodiversity species areas land protected

Topic	Human-labelled keywords	Model keywords	tf-idf keywords
349	tannins and polyphenols	hydrolyzable ellagitannins tannins phenolics polyphenol tannin polyphenols dad phenolic hplc condensed extracts flavonoids esi medicinal uplc glycosides antioxidant quercetin flavonol leaves chromatography fruits ionization_mass metabolites	tannins plant phenolic acid activity ellagitannins

Appendix F. Abbreviations of organisations

Abbreviation	Name in English	Name in Finnish
Aalto	Aalto University	Aalto-yliopisto
FFA	Finnish Food Authority	Ruokavirasto
FIIA	Finnish Institute of International Affairs	Ulkopoliittinen instituutti
FIOH	Finnish Institute of Occupational Health	Työterveyslaitos
FMI	Finnish Meteorological Institute	Ilmatieteen laitos
Govt Res	All other government institutions, excluding actual research institutes such as VTT	Muut valtion laitokset poislukien varsinaiset tutkimuslaitokset
GTK	Geological Survey of Finland	Geologian tutkimuskeskus
HU Hosp	Helsinki University Central Hospital	Helsingin seudun yliopistollinen keskussairaala
JYU	University of Jyväskylä	Jyväskylän yliopisto
Luke	Natural Resources Institute Finland	Luonnonvarakeskus
LUT	LUT University	Lappeenrannan–Lahden teknillinen yliopisto LUT
NLS	National Land Survey of Finland	Maanmittauslaitos
OU Hosp	Oulu University Hospital	Oulun yliopistollinen sairaala
SHH	Hanken School of Economics	Svenska handelshögskolan
SHP	Hospitals of hospital districts, excluding university hospitals	Muut sairaanhoitopiirien sairaalat poislukien yliopistosairaalat
STUK	Radiation and Nuclear Safety Authority	Säteilyturvakeskus
SYKE	Finnish Environment Institute	Suomen ympäristökeskus
TAU	Tampere University	Tampereen yliopisto
TAU Hosp	Tampere University Hospital	Tampereen yliopistollinen sairaala
THL	Finnish Institute for Health and Welfare	Terveyden ja hyvinvoinnin laitos
UEF	University of Eastern Finland	Itä-Suomen yliopisto
UEF Hosp	Kuopio University Hospital	Kuopion yliopistollinen sairaala
UH	University of Helsinki	Helsingin yliopisto

ULA	University of Lapland	Lapin yliopisto
UNIARTS	University of the Arts Helsinki	Taideyliopisto
UO	University of Oulu	Oulun yliopisto
UTU	University of Turku	Turun yliopisto
UTU Hosp	Turku University Hospital	Turun yliopistollinen keskussairaala
UVA	University of Vaasa	Vaasan yliopisto
VATT	VATT Institute for Economic Research	Valtion taloudellinen tutkimuskeskus
VTT	VTT Technical Research Centre of Finland Ltd	Teknologian tutkimuskeskus VTT Oy
ÅAU	Åbo Akademi University	Åbo Akademi

Appendix G. Model parameter settings

The overall goal of topic models is to generate good, interpretable topics in which all publications can be grouped together using a single coherent concept (Mimno, et al., 2011). There are multiple metrics for evaluating topic models. However, optimizing such quantitative metrics do not necessarily lead to topics being easily interpretable by humans (Chang, et al., 2009). In this exploratory project, four coherence metrics, C_{UMASS}, C_V, C_{UCI}, and C_{NPMI} (Röder, et al., 2015) as implemented in gensim (Řehůřek, 2021), were used to guide human judgment on the number of topics and other hyperparameters.

Latent Dirichlet Allocation model

The following parameters were selected for the presented latent feature topic model:

- number of topics: 1026
- parameter of the prior topic distribution of publications α: 0.1
- parameter of the prior word distribution of topics β: 0.01
- probability of a word being generated by the latent feature model λ: 1.0 (in other words, the model uses only the latent feature model)

Word embeddings in doc2vec

The following parameters were used when creating word embeddings using the gensim implementation (Řehůřek, 2021):

- distributed bag of words (PV-DBOW) (dm: 0)
- training of word-vectors simultaneously with DBOW document vectors (dbow_words: 1)
- dimensionality of vectors: 300 (vector_size: 300)
- window: 15
- threshold for downsampling higher-frequency words (sample: 1e-5)
- epochs: 400
- hierarchical soft max (hs: 1)

The other parameters were set to defaults. These settings were guided by the recommendations by Lau and Baldwin (2016).