Issue 7 / Winter 2021/2022

SOBIGDATA RESEARCH INFRASTRUCTURE

Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Magazine

Editorial

Trust with reference to the blooming digitalization is becoming a buzzword.
It is naturally related to concepts such as transparency, accountability, and explainability.

Many policy documents, at least at the European Union level, leverage on the notion as a flagship one for embedding values in developing technologies related to the digitalization of economies and societies.

After all, to develop and maintain trust information need to be transparent and understandable to their addressees. Their end users will need to be able to decide upon that information, relying on the various players and on their responsibility. There is no Trust without accountability; there is no accountability without transparency.
In the digital economy many technological and regulatory layers overlap and interact creating a complex puzzle of connected digital ecosystems.

In the face of rapidly evolving technologies, regulation

[Continues on pag. 3]

Inside this issue

Editorial

News

Events Highlights

Call for Paper

TransNational Access

Research Highlights

Exploratories Highlights

Trust with reference to the blooming digitalization is becoming a buzzword

Giovanni Comandé, LiderLab, Scuola Superiore Sant’Anna, Pisa

and ethical constraints have often been accused of being an obstacle to innovation. On the contrary, the axiological dimension developed by the EU in recent years has made such ethical dimension faithful to its founding values its flag and compass. It has connected ethics, law and technology. A clear example is reflected in the Guidelines for trustworthy Ai developed by the EU Commission’s High Level Group.

> “The more powerful and disruptive the technologies are, the more the respect for the core values and principles of ethics need to be at their heart. ”
>
> *Giovanni Comandé, LiderLab Scuola Superiore Sant’Anna*

Sometimes, the ethical dimension has begun to be reflected operationally in regulatory prescriptions. It is certainly the case of GDPR where the axiological dimension, based on the protection of fundamental rights has been translated operationally in the implementation of the principles of accountability and of transparency in its basic rules.

Since its inception (art. 1.1) the GDPR clarifies that it “lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data” (emphasis added). Protecting fundamental rights is conceived structurally as not antithetic to the free movement of personal data. This is a key point since it sets the need for a continuous balancing between the mandatory respect for fundamental rights and the need to circulate personal data. After all, the richness of data lies ”in its use and re-use”.

The overarching idea stressed in this first white paper of the High Level Advisory Board of SoBigData++ is that the goal to make data «As close as necessary, as open as possible» needs to be enabled by a clear framework respecting EU values and legal constraints.

The legal and ethical framework, as we start to demonstrate in this paper, is not a roadblock to innovation. To the contrary it enables it in a sustainable way. This is also the reason for which the emerging legal and ethical framework for leveraging the data society and for developing and deploying related AI systems is constantly referring -expressly or implicitly- to the 17 UN Sustainable Development Goals. Societies and mankind in general cannot afford anymore to overlook sustainability of technological developments on all grounds.

The more powerful and disruptive the technologies are, the more the respect for the core values and principles of ethics need to be at their heart.

The paper, as SoBigData++, naturally refers to innovations in the data domain. We claim that a really successful data science, capable to expand the accessibility of data for research and innovation is both bound and enabled by a clear legal and ethical framework.

Our examples of methods for facilitating data sharing, privacy-preserving technologies, decentralization, data altruism and their connection with existing and forthcoming regulations (e.g. the interplay between the Data Governance Act and the GDPR), illustrate how this fundamental understanding of the role of regulation as an enabler of research, innovation and data sharing reaches well beyond the EU borders and can set a benchmark for all industrial societies.

We will move further in this process of advocating the power of compliance for promoting innovation. More to come!

TransNational Access: a program of Short-Term Scientific Missions to carry forward your own big data project

We welcome applications from individuals with a scientific interest, professionals, startups and innovators that may benefit from training in data science and social media analytics.

The SoBigData++ RI manages vertical, thematic environments, called exploratories, on top of the SoBigData infrastructures, for performing cross-disciplinary social mining research. The Transnational Activities offered in this call will be for Short-Term Scientific Missions (STSM), between 3 weeks and 2 months.

Under this call, there will be two kinds of proposals funded: **STSM research proposals** and **STSM tool/data integration proposals.**

Funding is available **up to 5000 euros** per participant (to cover the cost of daily subsistence, accommodation, and economy flights/train).

STSM bursaries are awarded on a competitive basis, according to the procedure described in the application pack and eligibility criteria below, and based upon the quality of the applicant, the scientific merit of the proposed project, and their personal statement.

Applications from female scientists are particularly encouraged.

Visitors are welcome subject to the host country’s and host institution’s **Covid-19 regulations.** We will consider offering up to a 6 month postponement of an accepted application if travel restrictions are imposed due to Covid-19.

APPLY NOW!

Visit our website
http://www.sobigdata.eu/content/call-2021-22-sobigdata-transnational-access

Pre-requisites for projects to carry out hosted research:
- Good understanding of social data and, ideally, track record of prior social data analysis projects;
- Experience with using at least one of machine learning, natural language processing, and/or complex networks algorithms.

Pre-requisites for projects to integrate new tools/datasets/services:
- An already existing open-source tool for social media mining to be integrated or an already created openly licensed dataset of relevance to SoBigData++, that can be integrated within the infrastructure.

The goal is to provide researchers and professionals with **access to big data computing platforms, big social data resources, and cutting-edge computational methods.**

STSM visitors will be able to:
- Interact with the local experts
- Discuss research questions
- Run experiments on non-public big social datasets and algorithms
- Present results at workshops/seminars

The STSM visits will enable multi-disciplinary social mining experiments with the SoBigData++ Research Infrastructure assets: big data sets, analytical tools, services and skills.

Photo_credit_JoshuaWoroniecki - Pixabay

SoBigData++ and ACCORDION: extending Research Infrastructures at the Edge

An infrastructure offers means to define complex analytical processes and workflows, bridging the gap between experts and analytical technology.

Patrizio Dazzi, ISTI CNR / patrizio.dazzi@isti.cnr.it

Research infrastructures play a crucial role in the development of data science. In fact, the conjunction of data, infrastructures and analytical methods enable multidisciplinary scientists and innovators to extract knowledge and to make the knowledge and experiments reusable by the scientific community, providing an impact on science and society.

An infrastructure offers means to define complex analytical processes and workflows, thus bridging the gap between experts and analytical technology. Experiments in turn generate new relevant data, methods, and workflows that can be integrated into the platform by scientists, contributing to the expansion of the RI itself. As a matter of fact, the availability of data creates opportunities but also new risks. The use of data science techniques could expose sensitive traits of individual persons and invade their privacy.

On the other hand, Edge computing is a novel computing paradigm that is spreading and developing at an incredible pace. Edge computing is based on the assumption that for certain applications is beneficial to bring the computation, and keep data, as closer as possible to data or end-users.

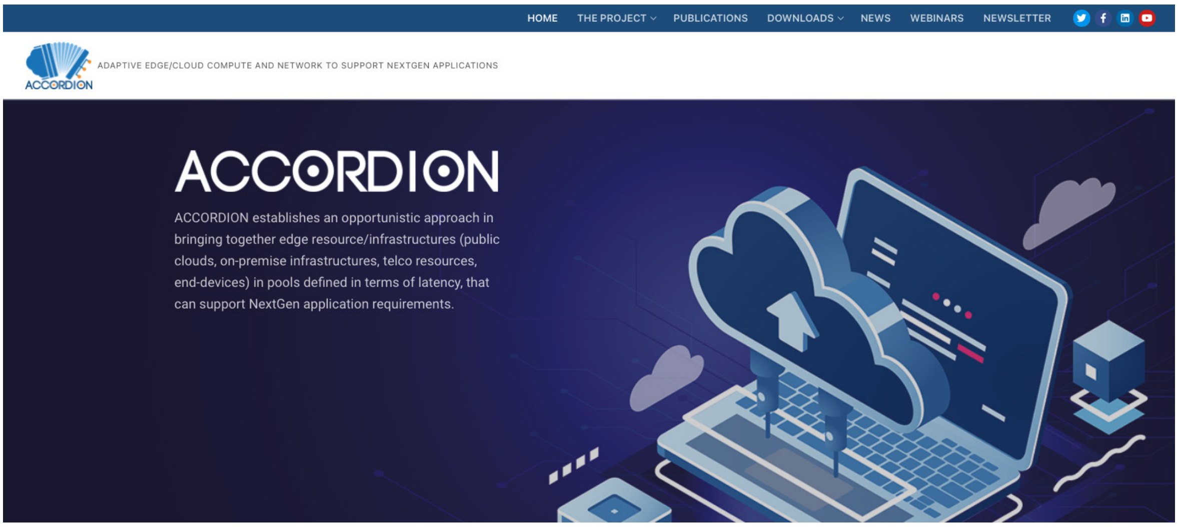
H2020 ACCORDION project establishes an opportunistic approach in bringing together edge resource/infrastructures (public clouds, on-premise infrastructures, telco resources, end-devices) in pools defined in terms of latency, that can support NextGen application requirements.

ACCORDION intelligently orchestrates applications on the compute & network continuum formed between edge and public clouds, using the latter as a capacitor. Deployment decisions will be taken also based on privacy, security, cost, time, and resource type criteria.

ACCORDION and SoBigData++ are discussing on a collaboration for supporting the extension of Research Infrastructures at the Edge. More specifically, the plan is on the exploitation and adaptation of the ACCORDION application model to support SoBigData++ workflows. A preliminary investigation, presented at the ACM FRAME Workshop 2021, has been published recently [1].

[1] Valerio Grossi, Roberto Trasarti, and Patrizio Dazzi. 2020. Data Science Workflows for the Cloud/Edge Computing Continuum. In Proceedings of the 1st Workshop on Flexible Resource and Application Management on the Edge (FRAME '21). Association for Computing Machinery, New York, NY, USA, 41–44. DOI:https://doi.org/10.1145/3452369.3463820

For more information, visit: https://www.accordion-project.eu/



ADAPTIVE EDGE/CLOUD COMPUTE AND NETWORK TO SUPPORT NEXTGEN APPLICATIONS

ACCORDION

ACCORDION establishes an opportunistic approach in bringing together edge resource/infrastructures (public clouds, on-premise infrastructures, telco resources, end-devices) in pools defined in terms of latency, that can support NextGen application requirements.

TU Delft and WHO workshop on Design for Values in healthcare

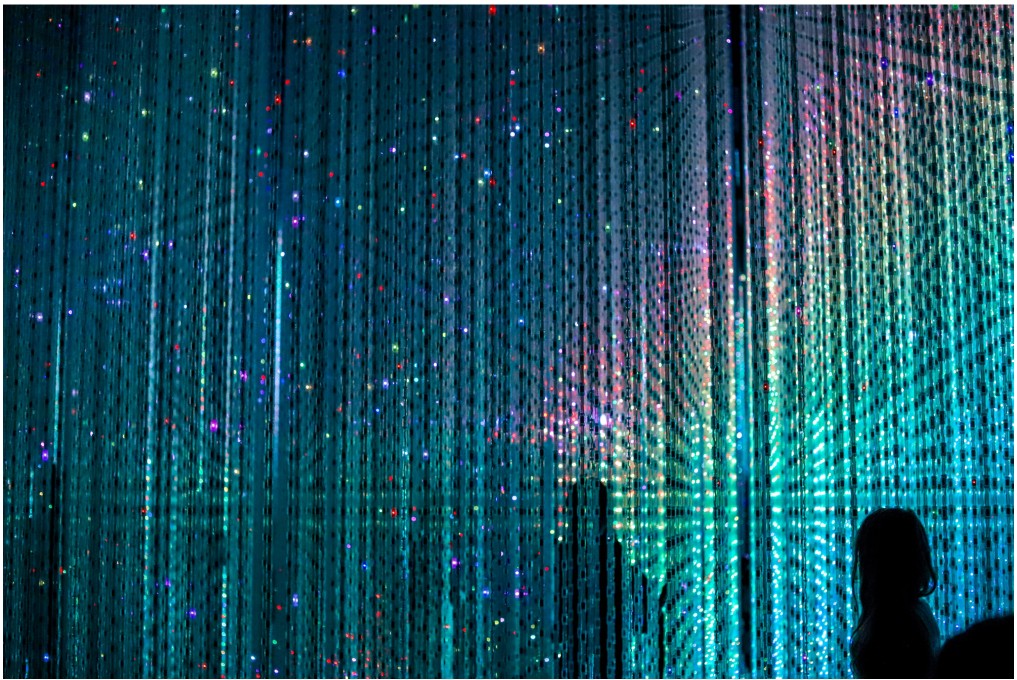
On November 12th 2021 the Delft University of Technology in collaboration with the World Health Organization hosted a workshop entitled “Ethics and Governance of Artificial Intelligence (AI) for Health: The Importance of Design for Values.” This collaboration aimed at pointing out the importance of designing values into artificial intelligence systems used in healthcare.

Giorgia Pozzi, Delft University of Technology | g.pozzi@tudelft.nl

Juan M. Durán, Delft University of Technology | j.m.duran@tudelft.nl

Jeroen van den Hoven, Delft University of Technology | j.m.vandenhoven@tudelft.nl

The development of artificial intelligence-based technologies to be introduced in the field of medicine and healthcare is increasing rapidly and holds great potential for the improvement of healthcare delivery and patient care. Along with the increasing use of AI systems in the context of medicine and healthcare, academic and public discussions regarding the ethical requirements that these systems must fulfill have come to the fore. Particularly important is the question of how to make ethical principles and human values to bear upon these technologies effectively or how to make them part of their design. In fact, if we want these systems to successfully become part of daily medical practice, we cannot compromise on the ethical and scientific standards they need to uphold.



Photo_Credit_ Mathis_Jrdl_Unsplash

The methodology of Design for Values, that has been developed and promoted by the TU Delft research community over the last decade can offer fruitful perspectives on how to (i) arrive at responsible innovations and (ii) guarantee the ethical use of AI technologies in the field of healthcare. This should happen through appropriate conceptual engineering of relevant value, concepts followed by requirements engineering, and eventually engineering design that build upon these.

There are for example many different conceptions of fairness that are used, implied or tacitly assumed when discussing ‘algorithmic fairness’. Which one do we choose, when and where? And why? These questions need to be answered before computer scientists and medical informatics experts can develop AI applications.

Similar questions can be raised regarding explainability, privacy, responsibility, accountability, sustainability and many more value concepts. Indeed, this conceptual analysis or conceptual engineering is needed in order to successfully introduce these systems into healthcare practices.

Against this background, the main goal of the workshop was to explore the potential and possible challenges of Design for Values approach in the context of medical AI. As such the workshop was one of the next steps of the WHO to turn Ethics Guidance into an actionable framework for global health care practitioners and clinical AI researchers and developers.

During the workshop, talks given by experts in the field were followed by fruitful exchanges with the audience and panel members:

Prof. Jeroen van den Hoven kickstarted the event with a presentation of the Delft approach on Design for Values and its significance, especially in high-stakes fields such as medicine and healthcare. In particular, he stressed the concrete role played by philosophical work and theoretical conceptualization in formulating clear requirements that these technologies should be endowed with. Moreover, he emphasized how conceptual clarity can be of particular importance for the daily work of engineers and computer scientists developing these systems to render these technologies ethically and socially acceptable.

Prof. Ibo van de Poel (TU Delft) addressed the question of how to translate values into design requirements, specifying the ways in which we can make sure that a particular value, that we deem important to be implemented into a technology, is indeed met by the ways in which the system is operated. That is to say, he addressed the important issue of how to make sure that an intended value is indeed the realized value when the technology, in our case medical AI systems, becomes part of a sociotechnical system.

Christian Quintero from the Universidad Militar Nueva Granada (Colombia) expanded on which approaches can be taken if we want to operationalize human values into technological systems.

At the center of the panel discussion were issues revolving around the question of how to render AI technologies inclusive and participatory. An especially salient question was also how to promote a participatory approach in middle- and low-income countries that have mostly a more constrained access to where the design of these systems are put forward.To foster the responsible development of AI in the health sector, the WHO issued a guidance for ethics and governance of artificial intelligence for health earlier this year [1].

Prof. Jennifer Gibson from the University of Toronto reported the main findings ensuing from the WHO ethical guidelines of AI systems in the healthcare domain. The aim of the WHO report is to clarify rules and principles that are considered to be particularly important for the ethical development and use of these systems that should effectively support the working physicians, healthcare personnel and patients, as well as other stakeholders involved in and affected by automated decision-making. Partnering up with the WHO in the organization of this workshop represents a first step to open a dialogue in a multi-disciplinary fashion and to promote an ethically aware co-design of AI technologies that can benefit the healthcare sector.

Links
[1] World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance (https://www.who.int/publications/i/item/9789240029200)

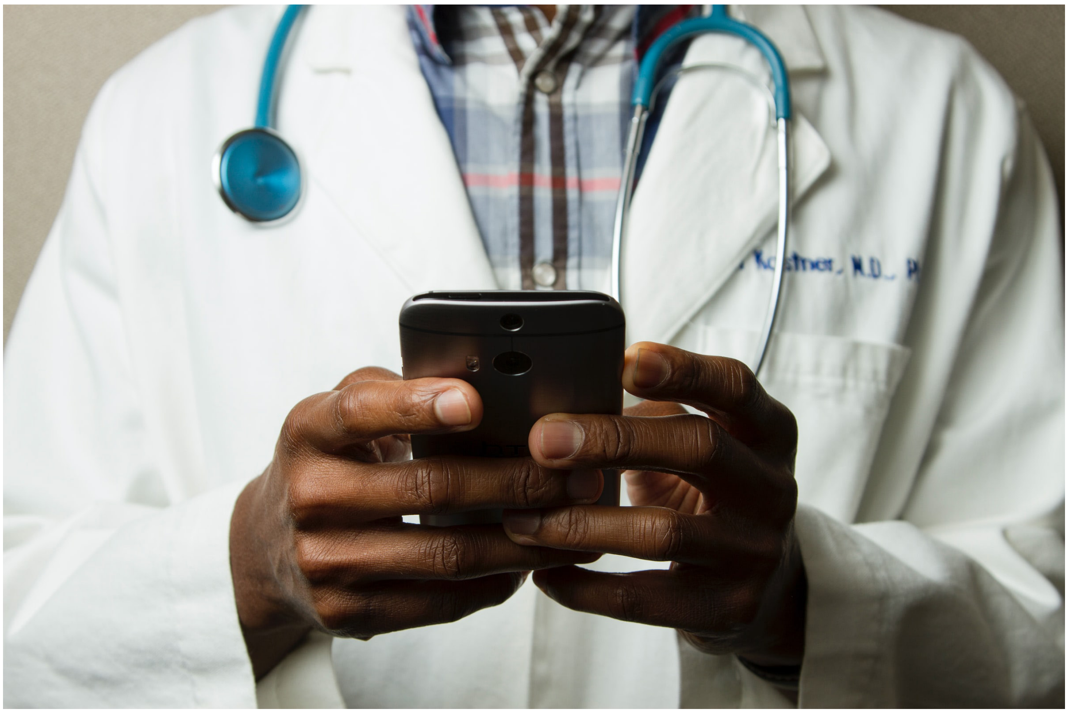


Photo credit: National Cancer Institute \ Unsplash

SoBigData **Events**

From large conferences to smaller webinars, multiple events have continued to take place in a hybrid mode this year.

The editorial Board

With the international travel situation continuing to be highly unpredictable and many countries still experiencing restrictions in some form or another, the SoBigData++ project has continued to plan, host and participate to events in a virtual format. There has been a variety of events in the past 6 months of 2021 – some of the highlights are detailed below.

AI & SOCIETY ROUNDTABLE

This roundtable took place on 30 June 2021. The roundtable was designed as a collective intelligence exercise towards shaping the research questions of Social AI, driven by societal challenges. It was implemented through a structured conversation among inter-disciplinary scientists, looking at the relationship between AI and society from multiple perspectives. [L1].

5TH SOBIGDATA++ AWARENESS PANEL:

Legal Materials as Big Data: (algo) Rithmsto Support Legal Interpretation. A Dialogue with Data Scientists [L2]

AUTOMATED METHODS OF URBAN GREEN ANALYSIS

A webinar aimed at providing a preliminary definition of the state-of-the-art upon automatic methods for systematized urban green data collection [L3]. The talk focused on the methods and tools that are currently available for the analysis of urban green, considering their degree of accuracy (e.g. location, size, aboveground volume, canopy cover, leaf area, species identity) in relation to the development of the urban green infrastructures.

The speaker was **Giorgio Vacchiano**, Researcher and Associate Professor of Forest Management and Planning at the Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy (DISAA) of the University of Milan.

SOBIGDATA @ DSAA: A RESEARCH INFRASTRUCTURE TO EMPOWER DATA SCIENCE ANALYSIS

This tutorial [L4] was part of DSAA 2021. The objectives of the tutorial was to show how SoBigData RI can support data scientists in doing cutting edge science and experiments.

The IEEE International Conference on Data Science and Advanced Analytics (DSAA) features its strong interdisciplinary synergy between statistics (sponsored by ASA), computing, and information/intelligence sciences (by IEEE and ACM), and cross-domain interactions between academia and business/industry for data science and analytics. DSAA sets up a high standard for its organizing committees, keynote speeches, submissions to the main and special session tracks, and a competitive rate for paper acceptance.

SOBIGDATA R.I. AT SCIDATACON 2021

The format of this session at SciDataCon2021 was a mix of research and practice presentations where the SoBigData++ project was presented in all its parts. [L5]

SciDataCon is the international conference for scrutiny and discussion of the frontier issues of data in research. The scope of SciDataCon covers policy matters and the place of data in the scientific endeavour and scholarly communications; the opportunities of the data revolution for the global research enterprise; innovations in data science and data stewardship; and the challenge of developing a sustainable data ecosystem, including the role of education and capacity building.

SOBIGDATA AT HUMANE-AI-NET

The format of this session workshop was a mix of research and practice presentations where the SoBigData++ project was presented in all its parts.

MOBIDATALAB WEBINAR

[6]Data sharing in the transport and mobility industries has begun to rise in recent years as stakeholders such as transport authorities, operators and other mobility actors try to address key issues and challenges through collaboration. Fostering such a culture is, thus, instrumental to have a more sustainable and interconnected Europe, based on quality, accessible and usable mobility data.

Links

[L1] http://www.sobigdata.eu/events/ai-society-roundtable

[L2] https://tinyurl.com/4rxjm9w7

[L3] https://tinyurl.com/398efcca

[L4] http://www.sobigdata.eu/events/sobigdataeu-research-infrastructure-empower-data-science-analysis

[L5] http://www.sobigdata.eu/events/sobigdata-ri-scidatacon-2021
[L6]http://www.sobigdata.eu/events/webinar-mobidatalab

PSD 2022: Privacy in Statistical Databases 2022

Paris, France, September 21-23, 2022
http://unescoprivacychair.urv.cat/psd2022
Submission deadline: *MAY 15, 2022*

CALL FOR PAPERS

Privacy in statistical databases is about finding tradeoffs to the tension between the increasing societal and economical demand for accurate information and the legal and ethical obligation to protect the privacy of individuals and enterprise which are the respondents providing the statistical data. In the case of statistical databases, the motivation for respondent privacy is one of survival: data collectors cannot expect to collect accurate information from individual or corporate respondents unless these feel the privacy of their responses is guaranteed.

Beyond respondent privacy, there are two additional privacy dimensions to be considered: privacy for the data owners (organizations owning or gathering the data, who would not like to share the data they have collected at great expense) and privacy for the users (those who submit queries to the database and would like their analyses to stay private).

“Privacy in Statistical Databases 2022” (PSD 2022) is a conference organized by the CRISES research group at Universitat Rovira i Virgili with proceedings published by Springer-Verlag in Lecture Notes in Computer Science. The purpose of PSD 2022 is to attract world-wide, high-level research in statistical database privacy.

PSD2020 is a successor to PSD 2020 (Tarragona, Sep. 23-25, 2020), PSD 2018 (València, Sep. 26-28, 2018), PSD 2016 (Dubrovnik, Sep. 14-16, 2016), PSD 2014 (Eivissa, Sep. 17-19, 2014), PSD 2012 (Palermo, Sep. 26-28, 2012), PSD 2010 (Corfu, Sep. 22-24, 2010), PSD 2008 (Istanbul, Sep. 24-26, 2008), PSD 2006 (Rome, Dec. 13-15, 2006) and PSD 2004 (Barcelona, June 9-11, 2004), all with proceedings published by Springer in LNCS 12276, LNCS 11126, LNCS 9867, LNCS 8744, LNCS 7556, LNCS 6344, LNCS 5262, LNCS 4302 and LNCS 3050, respectively. Those nine PSD conferences follow a tradition of high-quality technical conferences on SDC which started with “Statistical Data Protection-SDP’98”, held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published in Springer LNCS 2316.

Like the aforementioned preceding conferences, PSD 2022 originates in Europe, but wishes to stay a world-wide event in database privacy and SDC. Thus, contributions and attendees from overseas are welcome.

PRIVACY IN STATISTICAL DATABASES 2022
Paris, France. September 21-23, 2022

TransNational Access and the impact of Covid19

At the intersection between social sciences and statistical physics, our new methodological approach paves new ways for defining online collective identities in a fully data-driven fashion.

The Editorial Board

Covid-19 has obviously had a huge impact on the project resulting in the postponement of all TA visits. With the various local, regional, national and international travel restrictions the Call was held back and all TA activities were put on hold.

A close eye was kept on the ever-changing pandemic situation and when it was deemed viable the Call was released in mid-June 2021. The number of applications has been inhibited due to travel issues and it is expected this will continue to be the case while the pandemic persists.
The project has put a number of processes and checks into place to provide some flexibility to potential visitors. The host and applicant, once their application has been approved, will liaise and discuss the timings of a visit and the travel restrictions of both departure and destination country. The host will provide some flexibility in case of unforeseen or unexpected circumstances and the visitor will be advised to factor in the unforeseen and unexpected into their plans as Covid-19 related issues are outside of the project's control. Furthermore, there is now an option for a successful applicant to postpone their visit for up to 6 months which should provide sufficient flexibility to a visitor.

The project is mindful that the pandemic may still be a feature for some time to come and as such will continue to monitor the situation and offer flexibility due to Covid-19 issues for the foreseeable future.

The first call was released in mid-June 2021 offering transnational visits, to complete a short-term scientific mission (STSM) of between 2 weeks to 2 months with up to €5,000 available to cover travel and living expenses. The project can now offer a visit to one of 17 European infrastructures providing wide ranging and varied expertise.

During this 2-year period (which includes 18 months of Covid-19 restrictions) the project received 12 applications of which 3 could not progress due to Covid-19 restrictions. Of the 9 that progressed, 7 were male applicants and 2 were female. There were 5 EU applicants and 4 non-EU (India and America) applicants.

The visits that have taken place have been successful and well received with the visitor gaining valuable experience and support for their projects and establishing meaningful academic connections and working relationships within the community.

SoBigData++ will continue to be as flexible and as accommodating as possible to TA applicants and visitors and it is hoped that the next period of the project will receive an increased number of applications.

Photo_Credit_Skitterphoto_Pixabay

Towards a Digital Ecosystem of Trust: Ethical, Legal and Societal implications: the first SoBigData White Paper

The success of digital economy and data-driven data science depend on the availability of sufficient data resources. How these resources become available relies on a complex factor interplay.

Giorgia Pozzi, Delft University of Technology | g.pozzi@tudelft.nl

Iryna Lishchuk, Leibniz Universität Hannover | lishchuk@iri.uni-hannover.de

The European Union's vision of a digital ecosystem of trust focuses on fundamental values such as fairness, transparency, and accountability. Aside from the focus on innovation and novel technological solutions, this constitutes a relevant novelty in contrast with the corporate-regulated US model or the state controlled technological landscape of countries such as China and Russia. In the context of the ongoing Covid-19 pandemic, the need for a strong transnational and safe data processing environment has emerged with great clarity. The assessment of factors that shape the digital ecosystem of trust and dictate responsible data science is at the core of the first SoBigData White Paper [1]. The white paper elaborates on use-cases from research projects, incentives for data sharing, privacy-preserving technologies, decentralization, data altruism, and explores avenues for data sharing opened by the legal framework of European Commission's Data Governance Act and the GDPR.

The European Union's willingness to propel technological development without compromising on legal and ethical standards (associated with the design, development, and implementation of new technologies) lead to a challenge: How are these two factors compatible? How can progress happen despite the barriers placed by the highest legal and ethical principles and requirements?

The answer lies in seeing ethics and legality as indispensable components of technological advancement.

Ethical and legal principles can coexist and facilitate socially oriented technological progress, without constituting an obstacle . Strategies that actively promote this approach have also been implemented within the SoBigData++ project. For example, exploratory Sustainable Cities for Citizens operates with mobility data collected from mobile phones, geolocated content uploaded to social media, etc.

Methods developed within the project suggest a possible and adequate response to potential intrusions into the private sphere (such as geo-location) of data subjects. These are, for example, privacy-by-design methods (Andrienko 2016) and also privacy risk estimators aimed to support data scientists in monitoring the risk of re-identification of individual mobility patterns and mobility profiles [4, 5].

Moreover, the development of new privacy-respecting technologies has been promoted within the SoBigData project. These approaches point at decentralization and incentivization [3] as useful ways to empower individuals, thus taking over control over the data sharing from third parties. Crafting decentralized protocols proves functional not only to achieve privacy by design but, arguably, also for the implementation of any ethical value into technology design.

Furthermore, the web's decentralization and de-monopolization of data are promoted by European regulatory and legal frameworks. In this respect, the Data Governance Act (DGA) aims to create favorable conditions in support of data sharing and altruistic uses of personal and non-personal data. For instance, the decentralized approach enables the sharing of sensitive categories of data selectively and in a privacy-preserving way. Due to high level of data quality and diversity, the European Union is particularly well-positioned to promote this approach. Moreover, this type of data sharing is reinforced by FAIR (Findability, Accessibility, Interoperability, and Reusability) principles dictating conditions for re-use which fall in line with the legal framework defined by the DGA and the GDPR.

The current Covid-19 pandemic highlighted the need of health data secondary processing, with pre-conditions such as stringent ethical and legal requirements.

In fact, both GDPR and normative codes of research ethics (such as the World Medical Association's Declaration of Helsinki) subject the processing of personal data concerning health for research to explicit consent by the data subject being informed about the essential points of research, approval by the ethics committee, the possibility to withdraw consent at any time, and appropriate data protection safeguards (e.g., de-identification, encryption, etc.). Such requirements are meant to ensure that research participants have some agency over the processing of their personal data.

In summary, there are numerous initiatives that contribute to the creation of an ecosystem of trust while following high ethical and legal values shared by European countries. Approaches that promote an ethics-by-design development of new technologies, privacy-preserving technologies, the phenomenon of data altruism, and strong legal frameworks are the key elements in furthering progress in reflection of the ethical and legal principles.

Selected References:
[1] Jeroen van den Hoven, et al., Towards a Digital Ecosystem of Trust: Ethical, Legal and Societal Implications, accepted for publication in Opinio Juris in Comparatione (www.opiniojurisincomparatione.org), November –December 2021
[2] Andrienko N. V., Andrienko G. L., Fuchs G. and Jankowski P. (2016). Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. Inf. Vis. 15(2): 117-153
[3] Domingo-Ferrer J. and Blanco-Justicia A. (2020). Ethical value-centric cybersecurity: a methodology based on a value graph. Science and Engineering Ethics, 26(3):1267-1285.
[4] Pellungrini R., Pappalardo L., Pratesi F. and Monreale A. (2018). A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. ACM Trans. Intell. Syst. Technol. 9(3): 31:1-31:27
[5] Pratesi F., Gabrielli L., Cintia P., Monreale A. and Giannotti F. (2020). PRIMULE: Privacy risk mitigation for user profiles. Data Knowl. Eng. 125: 101786

Photo_Credit_Ian_Battaglia_Unsplash

A dive into D4Science Infrastructure developments to master the opportunities offered by the platform enacting the implementation of the SoBigData ecosystem.

Massimiliano Assante, CNR-ISTI | massimiliano.assante@isti.cnr.it

Leonardo Candela, CNR-ISTI | leonardo.candela@isti.cnr.it

Pasquale Pagano, CNR-ISTI | pasquale.pagano@isti.cnr.it

The D4Science Infrastructure [R1] is the nucleus of the SoBigData ecosystem. Since the beginning it has been the platform enacting the development of the distributed and participatory digital infrastructure. During the last two years, its service offering was further extended to better match the needs and expectations of various communities of practice including SoBigData. D4Science is an IT infrastructure specifically conceived to support the development and operation of VREs by the as-a-Service provisioning mode. D4Science-based VREs are web-based, community-oriented, collaborative, user-friendly, open-science-enabler working environments for scientists and practitioners willing to work together to perform a certain (research) task. From the end-user perspective, each VRE manifests in a unifying web application (and a set of Application Programming Interfaces) (a) comprising several components made available by portlets organized in custom pages and menu items and (b) running in a plain web browser.

Every component is aiming at providing VRE users with facilities implemented by relying on one or more services possibly provisioned by diverse providers. In fact, every VRE is conceived to play the role of a gateway giving seamless access to the datasets and services of interest for the designated community while hiding the diversities originating from the multiplicity of resource providers. Among specific components each VRE offers, there are some basic elements enabling VRE users to perform their tasks collaboratively, namely: a workspace component to organise and share any digital artefact of interest; a social networking component to communicate with co-workers by posts and replies; a data analytics platform to share and execute analytics methods; a catalogue component to document and publish any worth sharing digital artifact. Fig. 1 depicts the service-oriented view of the D4Science architecture (the details are discussed elsewhere [R1]). Services are conceptually organized into three groups: front-end components called to realize the D4Science part user interacts directly; back-end components called to implement the business logic of D4Science; provided resources called to provide front-end components and back-end components with resources (computing, storage, data, software) to use. In order to complement this offering and bring community resources (the boxes with white names in Fig. 1) into VREs, the following three integration patterns are supported (besides implementing completely new services):
integrating existing applications;
integrating analytics methods and workflows;
integrating datasets and other resources for discovery and access.

D4Science offers four options for integrating existing applications (i.e. stand-alone systems offering one or more functionality either via web-based User Interfaces or via APIs) into VREs: (i) Adaption level integration where there is the willingness to fully integrate an existing application into a VRE, (ii) Adoption level integration where there is the willingness to integrate an existing application only for what regards its operation and management, (iii) Entry level integration where there is the willingness to reach a basic level of integration between the application and the VRE, (iv) Client integration where the willingness to integrate the application is unidirectional, meaning that the application will not be part of the VRE yet using the application it is possible to have access to the VRE resource space and contribute to it. When integrating an application, it is worth taking into account the security settings for D4Science and its VREs: (i) the communication between services hosted on diverse sites is counting on the Transport Layer Security protocol; (ii) the use of any service is regulated by authentication and authorization for both human users and applications; (iii) authorization is realized by a token-based approach where the token is associated with every interaction and used to verify whether the owner of the token is authorized or not to execute the action on the target resource; (iv) the authorization service exposes OAuth2 protocol APIs to enact any third-party application to interact with D4Science services, possibly on a user's behalf. Concerning the provisioning of the applications, this can be done either by the as-a-Service delivery model where the application runs on application owner premises or by the software package delivery model, including containerized applications to be hosted on D4Science computing infrastructure. Applications provisioned with the second approach include Docker containers to be deployed into a Docker Swarm cluster and Shiny Apps deployed on request by a ShinyProxy.

Regarding the integration of analytics methods, D4Science is equipped with a feature-rich platform for data analytics named DataMiner [R1]. This platform offers a web-based method integration environment for communities willing to transform almost any existing method and implemented procedure into an executable process offered by the platform. Whenever an existing method or procedure is integrated into the platform, (i) the method becomes an asset of the overall infrastructure, and it can be added to every VRE (if the license selected by the provider allows it); (ii) the method is transparently executed by relying on the D4Science distributed computing infrastructure designed to scale horizontally; (iii) a per-process graphical user interface is automatically generated thus to facilitate the execution (e.g. the compilation of parameters) and monitoring of the process; (iv) a standard API based on the WPS protocol is automatically generated, thus making it possible to programmatically invoke the process from existing clients; (v) a complete recording of every execution is automatically stored into the user workspace, including a provenance record enacting the repeatability of the specific computation; (vi) the method is published into the catalogue with rich metadata facilitating its discovery and use. D4Science complements this by offering Jupyter and RStudio environments as a service when requested by a VRE. In addition to that, D4Science offers the possibility to execute Java-based applications via SmartExecutor, either on request or by specific scheduling plan.

Finally, D4Science provides its users with a highly customisable catalogue offering its content via a GUI, a RESTful API (gCat) and some standards (e.g., DCAT, OAI-PMH) [R1]. This service is a key component of almost every VRE because it makes it possible for each community to organize a shared and searchable "research objects" space. Such a space is expected to be populated with descriptive records of any worth-sharing artifact that may or may not pre-exist the VRE. Research objects might represent datasets, software, services, processes, and the like.

These patterns have been widely used to develop the SoBigData infrastructure [L1]. Apart from the use of the catalogue [L2] to publish the great variety of its products (methods, datasets, training material) and supporting the development of the Literacy working environment, existing applications have been added to the platform by the app integration patterns as well as by analytics methods integration. Applications like TagMe, WAT and SWAT are now joined to D4Science by the adoption pattern. QuickRank and GATECloud have been integrated by developing specific analytics methods. In particular, in the case of GATECloud the analytics methods are simple wrappers taking care of invoking the homologous method on that platform. Twitter Monitor, an application to collect relevant messages from Twitter and share these with the SoBigData community make use of both Data Miner and SmartExecutor facilities to run on D4Science. The D4Science platform proved to be a key enabling technology for the development of the SoBigData infrastructure.

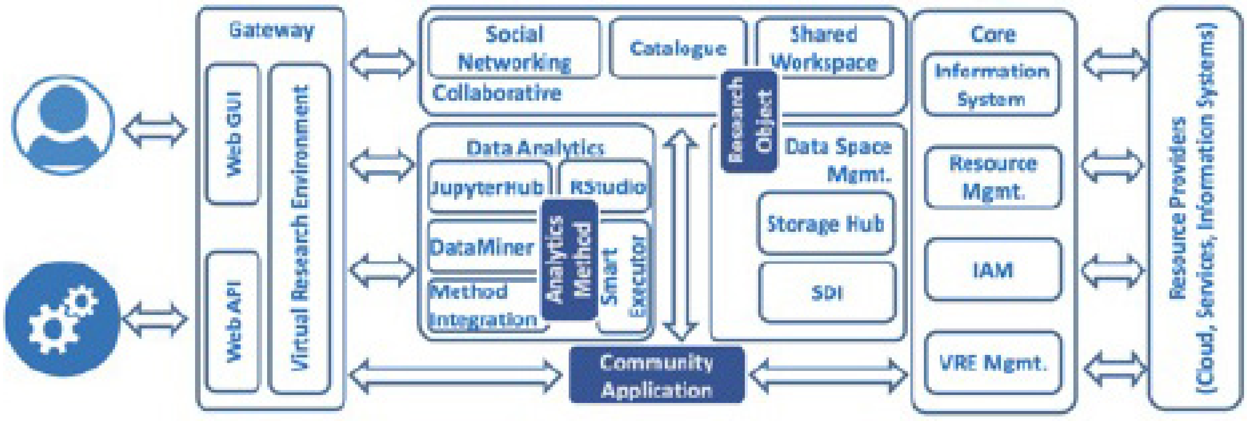


Fig. 1. D4Science Overall Architecture

Links:
[L1]: SoBigData Gateway https://sobigdata.d4science.org/
[L2]: SoBigData Catalogue https://sobigdata.d4science.org/catalogue-sobigdata
(put the URLs here and include in-text references)

References:
[R1]: M. Assante et al. (2019) Enacting open science by D4Science. Future Gener. Comput. Syst. 101: 555-563 https://doi.org/10.1016/j.future.2019.05.063

Urban green for Sustainable Cities for Citizens: new collaborations and insights on the SBD platform

Our cities stand today as a miraculous lab to foster an effective joint effort of different disciplines united by the common goal of setting our cities on a path towards well-being and social and environmental sustainability.

Simona Re, Eliante

Angelo facchini, IMT Lucca| angelo.facchini@imtlucca.it

Michela Natilli, CNR | michela.natilli@isti.cnr.it

Our cities stand today as a miraculous lab to foster an effective joint effort of different disciplines united by the common goal of setting our cities on a path towards well-being and social and environmental sustainability. As set out by the United Nations Sustainable Development Goals [1], the UNFCCC Paris Agreement [2] and main current European policies [3,4], cities of the world need to take urgent action against climate change and to better manage future exponential urban population growth. That means to effectively reinvent our lifestyles and cities in the direction of sustainability. Towards an effective and holistic urban development, data infrastructures including earth and in-field observations, mobility, health and social media data are key means and need to be better integrated to ensure the advancement of sustainability, biodiversity and climate change science in urban and peri-urban areas.

> «There is a need for research to rise to the challenge of monitoring and planning urban forests and their benefits in a changing climate»
>
> *Giorgio Vacchiano, University of Milan*

In particular, growing attention is given today to the study of urban green. Why studying trees in our cities is so important? For many reasons, given by the several benefits that urban trees provide, including cooling air, filtering air pollutants, regulating water flow, mitigate CO2 emissions, improving citizens health and well-being, saving energy for air conditioning and heating, and increasing urban biodiversity by providing food and protection to many species [5]. Taking into account the current effort of data scientists in the analysis of variables such as urban mobility, energy and quality of life, the related interconnections with urban green data open many insights for the development of new integrated methods and data-based approaches to urban metabolism analysis. In order to optimize the research effort and related outputs in this broad study horizon, the comparability and scalability of results from innovative tools are among the biggest challenges for the future research on cities.

Starting from this premise, the Institute of Information Science and Technologies of the Italian National Research Council (ISTI-CNR), IMT School for Advanced Studies Lucca and Eliante charity are currently collaborating to develop cutting-edge solutions for urban green assessment. Their results will support the scientific research on tools and methods for systematized urban green data collection and analysis, and may represent a precious starting point for new interesting scientific collaborations.

In particular, researchers from the Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy of the University of Milan are currently supporting the Sustainable Cities for Citizens task to reason and cooperate on further evolution and development of urban green studies on the SBD platform. The first collaboration is taking place in the context of the “Automated methods of urban green analysis - State of the art” micro-project, aiming at providing a preliminary definition of the state-of-the-art upon automatic and non automatic methods for systematized urban green data collection. In this regard, preliminary results of the micro-project activities have been discussed on July 13, 2021 by Giorgio Vacchiano, Associate Professor in Forest Management and Silviculture at the University of Milan, in the dedicated webinar entitled “Automated methods of urban green analysis”. «There is a need for research to rise to the challenge of monitoring and planning urban forests and their benefits in a changing climate» said Giorgio

Vacchiano. «We have summarized the most recent technological advances to support the assessment of urban forests, from remote sensing to simulation models of tree growth. We also highlighted which cities or regions with available open data about their urban tree - an example to be imitated by all administrations around the globe, with the support of technology and citizen science».

In line with the objectives and expectations of the project, hopefully new partners would get involved to nurture fruitful new research partnerships and collaborations for sustainable cities and urban green research in SBD. «We hope this is just the first of many collaborations to come. The increasing complexity of new societal and scientific challenges requires interdisciplinary research groups and a dialogue between different disciplines, backgrounds, and ways of thinking» said Luca Pappalardo, data scientist at KDD Lab of ISTI-CNR.

Visualizing the Results of Biclustering and Boolean Matrix Factorization Algorithms

Thibault Marette, Research Internship chez Kungliga Tekniska högskolan, Sweden

In many data science tasks our goal is to understand the relationship of two different sets of entities. For example, a common problem that online shops need to solve is to understand the relationship between customers and products. This relationship reveals clusters of products that are frequently bought together and groups of users who buy similar products. Such insights can then be used for marketing activities or to make product recommendations.

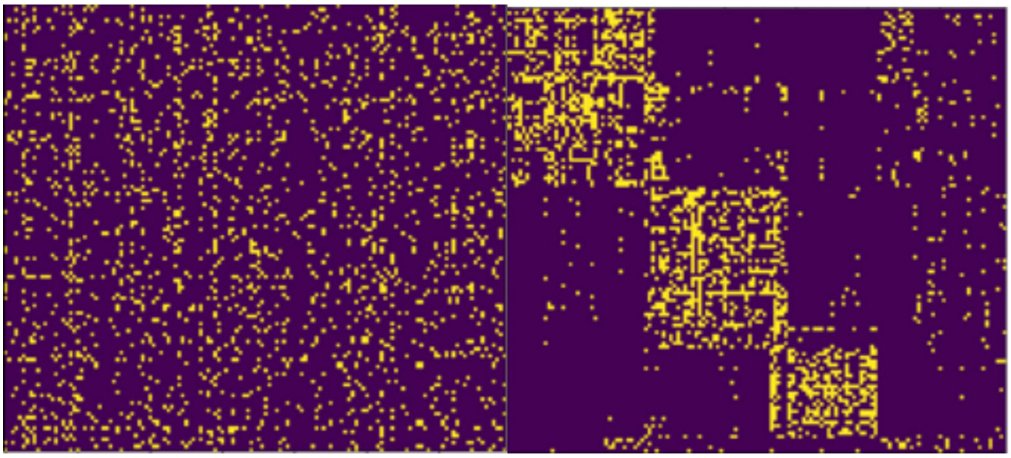


Figure 1. Visualization of an unordered matrix (left) and the same matrix after running a clustering algorithm and reordering the rows and columns based on the clustering (right).

This problem is often modelled using bipartite graphs where the two sides of the bipartite graph correspond to the two sets of different entities. An edge indicates that two entities interact with each other (e.g., a customer has bought a certain product). Then the goal is to find dense areas or clusters in this graph since the clusters reveal groups of entities that are related. This problem has applications in numerous fields [1][2][3].

Since nowadays the input datasets are very large, we cannot manually find such clusters. Therefore, automatic clustering algorithms were conceived and have been an active research area for decades [4][5][6]. However, using these algorithms can have drawbacks. First, the computed clusters are typically returned as a list of entities. This makes the analysis by a human difficult, since these lists do not reveal how strong the interactions are. Next, many of these algorithms need a set of parameters and often it is not clear how these parameters should be set for the data at hand. Finally, the computed clusters might overlap with each other by sharing entities, which complicates the interpretation of the results.

To mitigate these drawbacks we develop novel visualization tools and algorithms for drawing the outputs of clustering algorithms. Thereby we make the output of clustering algorithms more accessible and enable practitioners to use domain knowledge to assert the validity of a given clustering.

In our visualization we decide to draw the (bi-)adjacency matrix of the bipartite graph. Since the clusters represent dense subgraphs within the bipartite graph, each cluster will induce a dense area of 1-entries in the adjacency matrix as can be seen in Figure 1.

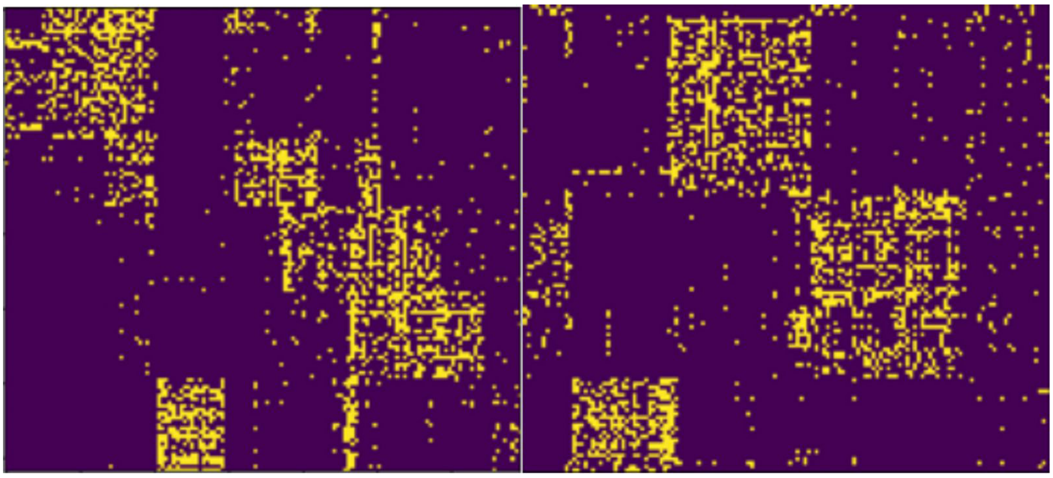


Figure 2. Visualization of two different clusterings of the same dataset. The picture visualizes the clustering results obtained from the PCV algorithm [5] (left) and the Basso algorithm [6] (right).

Therefore, analysts will be able to quickly understand the cluster structure of their data.

We propose an algorithm that, given a bipartite graph and a clustering of the graph, finds a suitable visualization of the adjacency matrix of the graph. Our approach is different from previous methods, which are unable to visualize a given graph clustering.

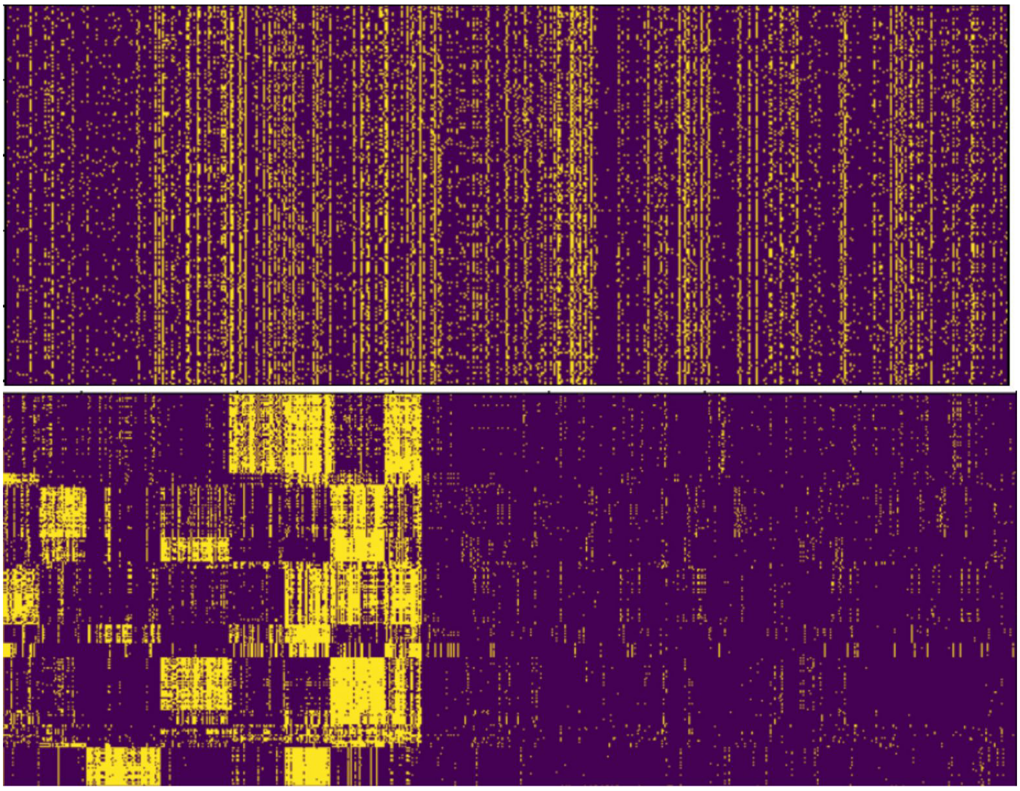


Figure 3. Visualization of a dataset before reordering (top) and after reordering (bottom). The dataset represents Finnish dialects and the regions they appear in. The visualization enables us to detect sparse and dense regions in the clustering.

To obtain a good visualization of the clustering we need to reorder the rows and the columns of the adjacency matrix based on the clustering. The choice of this ordering will have a large impact on how the clusters appear in the picture. Our main difficulty will be when the clusters returned by the clustering algorithm overlap, i.e., when the same entity appears in multiple clusters. To deal with this issue, we propose an objective function that generalizes to the overlapping setting and incentivizes orderings in which each cluster is drawn as a consecutive rectangle of 1-entries. We also provide algorithms to optimize this objective function.

By using our method, we obtain an intuitive way to compare the output of two different clustering algorithms. In Figure 2 we visualize the clusterings obtained from two different algorithms on the same dataset. The database used to obtain these figures is the Neogene of the Old World (NOW) database [3], which shows the relation between mammal fossil records and fossil localities in Europe.

Indeed, the left clustering in Figure 2 seems to describe many small clusters, whereas the clustering on the right of *Figure 2* describes larger clusters, with less noise outside of the clusters. Thus, it appears like the clustering on the left overfits the data and the clustering on the right fits the data better. These insights can be used by analysts to refine the clustering, either by manually changing the clustering or by adjusting the parameters of the algorithms to obtain better results.

Additionally, in *Figure 3* we present the visualization of a larger dataset before and after applying our algorithms. We can see that the dataset contains a strong cluster structure. Furthermore, we observe that some of the clusters are more dense than others, which can be important information for analysts to judge how strong the relations between the entities are.

In conclusion, we have presented algorithms that allow us to efficiently visualize the results of clustering algorithms. Our methods can be used to evaluate how each clustering shapes the data and it can enable analysts to assess how well a given clustering better describes the relationships within the data.

References

[1] Kemal Eren, Mehmet Deveci, Onur Küçüktunç, and Ümit V Çatalyürek. A comparativeanalysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics, 14(3):279–292, 2013.

[2] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray dataanalysis: from disarray to consolidation and consensus. Nature Reviews Genetics, 7(1):55–65, 2006.

[3] M Fortelius. New and Old Worlds Database of Fossil Mammals (NOW). 2015.

[4] John A. Hartigan. “Direct Clustering of a Data Matrix”. In: Journal of the American Statistical Association, 67.337 (1972), pp. 123–129.

[5] Stefan Neumann. Bipartite stochastic block models with tiny clusters. Conference on Neural Information Processing Systems, pp. 3871–3881, 2018.

[6] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. Thediscrete basis problem. IEEE Transactions on Knowledge and Data Engineering, 20(10):1348–1362, 2008.

Does physical activity change in immigrants within the European Economic Area? A case study of Italians move to Norway

Researchers have carried out systematic comparative analyses of migration and integration trends as well as analysed policies on migration and integration which are key as they not only influence migrants' ability to enter a country, but also the possibility to remain in the country, and their quality of life there.

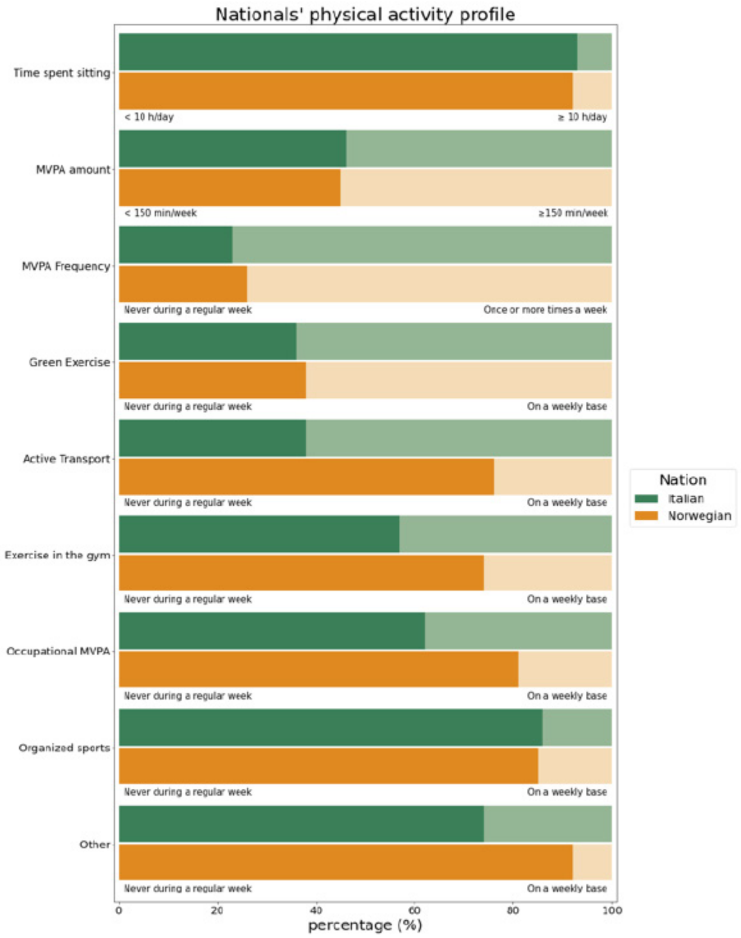
Alessio Rossi, University of Pisa, Italy

Physical activity (PA) is one of the lifestyle factors that most influence whether people live a long and healthy life. Not only does an active lifestyle make people live longer,but it also contributes to better mental health and, in general, a higher quality of life. In particular, the World Health Organization (WHO) recommends that, to improve and/or maintain good psycho-physical health, adults and elderly adults should engage in aerobic PA of light- or moderate-intensity (e.g., walking, cycling, swimming) for at least 150 min per week, or in aerobic PA of vigorous-intensity (e.g., running) for at least 75 min per week, or an equivalent combination of light-/moderate- and vigorous-intensity PA. The WHO also recommends performing exercises aimed at increasing muscular strength, flexibility and balance.

Moreover, it is recommended to avoid, for as much as possible, to spend prolonged time in inactivity or sedentary behaviors (e.g., sitting at work or watching TV). In spite of the consistent evidence on the health benefits of PA, as well as the clear and simple guidelines for PA behavior, insufficient PA remains one of the leading risk factors for poor health and mortality worldwide (1). From a health promotion perspective, a major challenge in promoting PA is that this behavior is subjected to social gradients, with more vulnerable sub-groups of the population less likely to engage in sufficient PA levels or respond to PA promotion initiatives. In particular, studies have consistently shown that gender, age, educational level, and ethnic background are major social determinants of PA behavior (2–4).

Compared to other western countries, Norwegians are estimated to be relatively active. Figures from the Global Health Observatory indicate that, in 2016, 68% of adult Norwegians met the WHO's recommended levels for PA (Figure 1). This prevalence was higher compared with other European countries such as, for example, Italy, where 59% of the adults engaged in sufficient PA levels. Moreover, compared with other countries, Norway shows a smaller gender-gap (Figure 1): while in Italy 54% of the men and 44% of women met the PA recommendations, in Norway 70% of men and 66% of women met the WHO's PA recommendations. Italian immigrants in Norway, although still relatively low in number, are a rapidly growing group. Italians enjoy the right of free movement of people within the European Union and the European Economic Area (EEA; which includes Norway). Italians' immigration to Norway has been steadily increasing since the establishment of the EEA Agreement in 1994, and it has 3-fold since the economic crisis of 2008 (5, 6). This trend is in line with the increased Italian mobility worldwide: from 2006 to 2019, the number of Italians registered as residents abroad increased by 70%, going from over 3.1 million to almost 5.3 million (5). These new waves of Italian immigrants have been often described as "the better youth" (young, well educated, cosmopolitan and mobile individuals), though there are also indications that grownups and families have been moving looking for jobs or better living conditions (6). According to figures from the Italian Embassy in Norway, to date 7.108 Italian citizens reside in Norway and are registered at the Norwegian Register of Italians Living Abroad (AIRE). Of these, 4.523 (2.862 men and 1.661 women) are Italian-born, while the others are progenies of Italian immigrants (6). In spite of this rapid increment, the living conditions of this group, especially in relation to their health and health-related behaviors such as PA, have received virtually no attention.

In a study of Calogiuri et al. (7) it was highlighted that a large majority of Italian immigrants in Norway perceived they were as or even more physically active in Norway than they would have been if they continued living in Italy. However, the prevalence of perceiving a negative impact was greater in specific subgroups (the men, older individuals, those who live in less urbanized regions of Norway, and those with lower socio-economic status). No significant differences between the Italian immigrants and the general Norwegian population were found for key indicators of PA levels, though some differences were observed in relation to specific activities. Associations of PA with different sociodemographic characteristics were observed, especially in relation to gender, educational level and, to a certain extent, age. In contrast with patterns observed in the general Italian population (as well as patterns observed in other immigration groups), women were generally found to be more physically active than men. These findings shed light on the PA habits of Italian immigrants living in Norway, a relatively small but rapidly growing immigration group and can be used to inform initiatives to promote PA in this or similar immigrations groups. This study indicates the potential of expanding the research on health and PA to under-researched immigrant groups, in particular within the EEA context. As mobility within the EEA is on the rise, it is important to understand how individuals interact with the opportunities and the culture of the country of resettlement, as well as how social gradients influence PA patterns in the context of migration.



Physical activity profile of Italian immigrants in Norway compared with the Norwegian general population

Exploratories: Migration Studies, Sport Data Science

Reference:

1. World Health Organization. WHO Guidelines on Physical Activity Sedentary Behaviour. Geneva:World Health Organization (2020).

2. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Health. (2006) 27:297–322. doi: 10.1146/annurev.publhealth.27.021405.102100

3. Trost SG, Owen N, Bauman AE, Sallis JF, Brown W. Correlates of adults' participation in physical activity: review and update. Med Sci Sports Exerc. (2002) 34:1996–2001. doi: 10.1097/00005768-200212000-00020

4. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? Lancet. (2012) 380:258–71. doi: 10.1016/S0140-6736(12)60735-1

5. Fondazione Migrantes. Rapporto Italiani nel Mondo 2019 [Report Italians in the world 2019]. TaAV editrice. (2019).

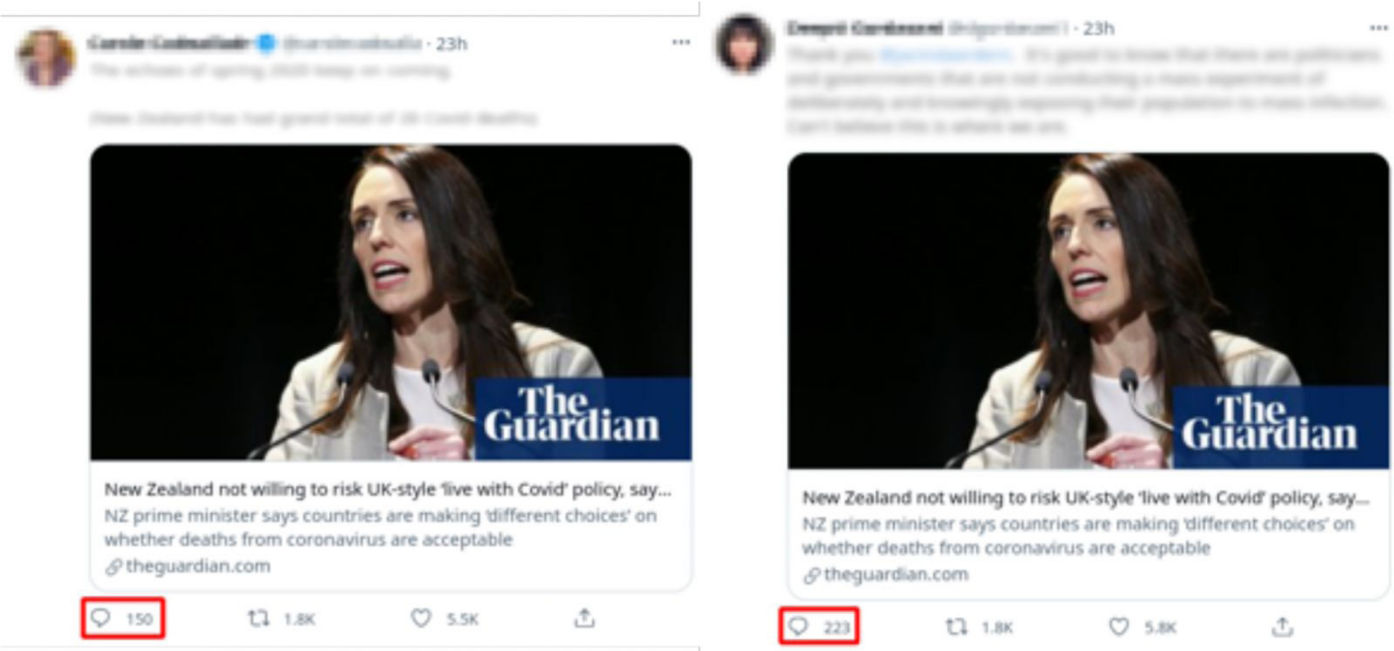
6. Miscali M, Calogiuri G, Terragni L. Bene, ma non benissimo: le nuove mobilità degli italiani in Norvegia ["Well, but not very well": new mobilities of Italian immigrants in Norway]. (2020).

7. Calogiuri G, Rossi A and Terragni L. Physical Activity Levels and Perceived Changes in the Context of Intra-EEA Migration: A Study on Italian Immigrants in Norway. Front Public Health. (2021) 9:689156. doi: 10.3389/fpubh.2021.689156

A Content-Based Approach for Detecting Echo Chambers in Online Social Networks

Online social networks are an integral part of modern society. Despite intensive research, many phenomena revolving around online social networks and their impact on society are still not sufficiently understood. One of the main challenges in this area is to establish whether echo chambers are widely present in social networks or not.

Francesco Zappia



[redacted] · 23h ···

[redacted]

New Zealand not willing to risk UK-style 'live with Covid' policy, say...
NZ prime minister says countries are making 'different choices' on whether deaths from coronavirus are acceptable
theguardian.com

150 1.8K 5.5K

[redacted] · 23h ···

[redacted]

New Zealand not willing to risk UK-style 'live with Covid' policy, say...
NZ prime minister says countries are making 'different choices' on whether deaths from coronavirus are acceptable
theguardian.com

223 1.8K 5.8K

Figure 1:Two threads on Twitter discussing a newspaper article. Both discussions give rise to a thread and both are about the same content (an article from The Guardian about New Zealand)

Online social networks are an integral part of modern society: they connect millions of users from around the world. For many people they serve as their main news aggregators and a window of what is happening in the world. Despite intensive research, many phenomena revolving around online social networks and their impact on society are still not sufficiently understood. One of the main challenges in this area is to establish whether echo chambers are widely present in social networks or not.

An echo chamber refers to a group of people who have the same opinions and reinforce their respective ideas without rebuttal from an opposing side [1]. Since it is believed that echo chambers play an important role in the online-radicalization of users, understanding whether echo chambers exist is an important question.

In our work, we introduce a new method for detecting echo chambers in online social networks. While previous approaches have mostly focussed on interactions between different users, our work also takes into account the contents associated to user interactions.

More concretely, consider the content given by a newspaper article. Our approach is to find all discussions about this article on online social networks, such as Twitter, and to download the corresponding (discussion) threads. While most newspaper articles will trigger only a small number of interactions between the users, controversial articles will trigger a heated discussion with lots of friendly and hostile interactions. In Figure 1 we present an example of two threads that discuss the same newspaper article and trigger many user interactions.

Therefore, our approach is to study the user interactions triggered by controversial newspaper articles. Our observation is that inside echo chambers, where beliefs are reinforced and where there is no rebuttal from outside, even controversial articles do not trigger hostile interactions. This suggests that if we can find local groups of users who agree on an article, even though that article is highly debated in the entire social network, then we can detect echo chambers.

As an example, consider the topic of politics in news media, which is known for being controversial [2]. Many articles about politics trigger a discussion between parties of opposing viewpoints, causing positive and negative interactions throughout the whole network. However, if a certain group of users discusses such controversial articles with mostly friendly interactions, then it is likely that all users in this group have a similar political stance and that they reinforce their opinions. Hence, such a group of users would constitute an echo chamber.

In our work we present a mathematical model for the problem of detecting content-based echo chambers. We then derive an integer linear program, which solves our problem exactly but does not scale to large datasets. To make our approach more scalable, we consider a relaxation of the integer linear program and use rounding heuristics to turn fractional solutions into integer-valued solutions. We also study the computational complexity of our problem and provide inapproximability results.

We evaluate our method on synthetic and on real-world data. On the synthetic data, our results show that our algorithm is able to reconstruct a set of planted echo chambers. We also run our heuristic algorithm on real-world data retrieved from Twitter and Reddit and observe that it finds subgraphs that match with our definition of content-based echo chambers.

References

[1] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, “Quantifying controversy on social media,” ACM Trans. Soc. Comput., vol. 1, no. 1, pp. 3:1–3:27, 2018. doi: 10.1145/3140565

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, “Political polarization on twitter,” in AAAI, vol. 5, no. 1, 2011.

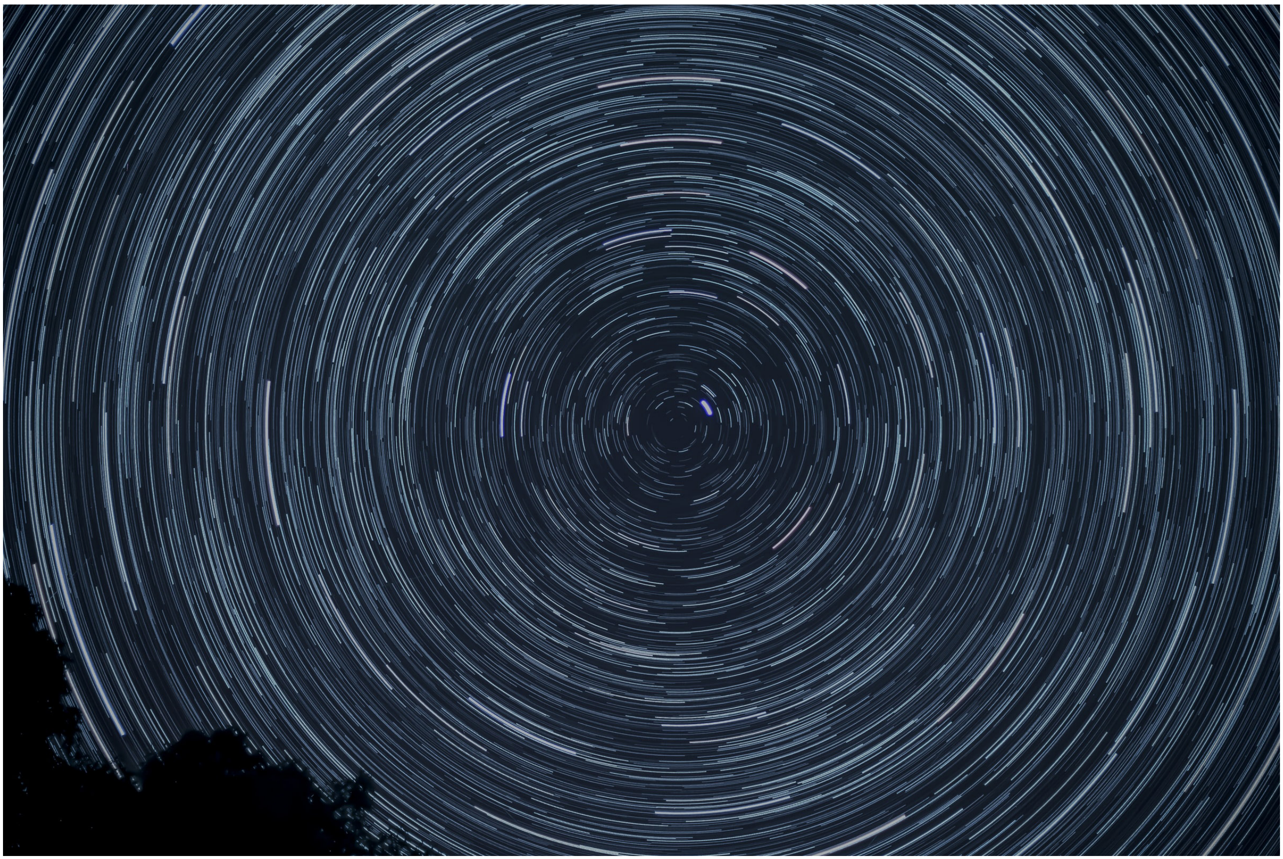


Photo credit: Patrick McManaman \ Unsplash

Academic Migration and Academic Networks: Evidence from Scholarly Big Data and the Iron Curtain

Big Data and the Iron Curtain are two phrases usually used to denote completely two different eras. Yet, the context the latter offers and the rich data source the former offers complement each other in a perfect manner enabling the causal identification of the effect of networks on migration.

Donia Kamel, Paris School of Economics | doniasameh3@gmail.com | Twitter: @Donia_Sameh

Cold War
Military alliances
Founding members of the North Atlantic Alliance (NATO) 1949
Entry: Greece and Turkey 1952, West Germany 1955, Spain 1982
Founding members of the Warsaw Pact 1955
Entry: East Germany 1956
Withdrawal: Albania 1968
Iceland
Sweden
Finland
Norway
Ireland
United Kingdom
Denmark
Nether-lands
Belgium
West Germany
East Germany
Poland
Soviet Union
Luxembourg
Czecho-slovakia
France
Switzer-land
Liecht.
Austria
Hungary
Romania
Yugoslavia
Monaco
San Marino
Italy
Vatican City
Andorra
Portugal
Spain
Bulgaria
Greece
Turkey
Cyprus
Morocco
Algeria
Tunisia
Malta

SO DO WE REALLY CARE?

The decision to migrate is one of the most important decisions an individual can make. As such, this decision is influenced and shaped by a lot of factors such as inequality levels at home and intended destination, returns to education, migration costs, employment prospects, life cycle considerations, and many more. Networks play a vital role in such decision as they influence all of the above factors in one way or another. Networks help the individual in learning about opportunities and conditions in potential destinations; moreover, at home and by construction, the structure of migrants' social networks shape their ability and desire to learn, and thus their migration prospects. On the macro level, studying migration of scientists is important as it has implications on brain drain; indeed, human capital influences differentials in economic prosperity across space and it is the engine of innovations and a major source of knowledge externalities. Thus, with the use of big data, the study of the role of networks in migration is facilitated, and this is the major contribution of this research as it is in the context of scientists.

HOW DOES THIS CONTRIBUTE TO THE LITERATURE?

There are ambiguities in the literature regarding the causal relationship between migration and network. This is due to the fact that it has historically been difficult to differentiate between distinct sources of social capital (synonymous to different types and structures of networks) in a single empirical setting. More specifically, in the migration case, traditional data sources inhibit the linking of social network structure to migration decisions. Additionally, the existing empirical evidence on the effects of networks makes the implicit assumption - which is a result of the constraints of the data - that all potential migrants benefit from the networks at destination equally (Bertoli and Ruyssen, 2018). This empirical evidence studying the effect of networks ranges from looking at share of households with a migrant at the village (McKenzie and Rapoport, 2010), size of diaspora at each destination country (Bertoli and Moraga, 2015 and Beine et al., 2011, 2015) or at the country level (Bertoli, 2010); whilst all making the implicit assumption that migrants benefit equally from networks.

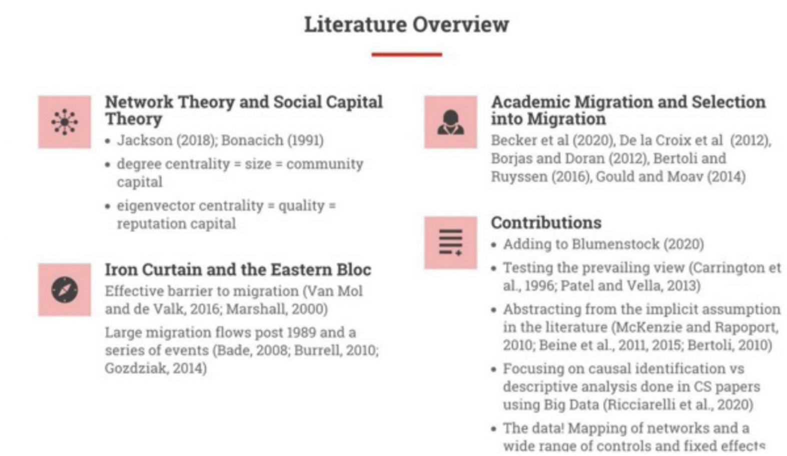
This research, instead, abstracts from this assumption by the ability to map and identify networks of individuals and specific characteristics about them, as also Blumenstock et al. (2020) recently did. Lastly and most importantly, this work attempts to reach causal identification of the effect of networks on migration decisions, by looking at a specific context in which manipulation of networks prior to migration was not possible and a rich data source that allows for a wide range of controls.

IDENTIFICATION OF THE CAUSAL EFFECT

Focusing on academics from Eastern Europe (henceforth EE) from 1980-1988 and their academic networks (1980-1988), I investigate the effect of academic network characteristics, by location, on the probability to migrate after the fall of the Berlin Wall in 1989 and up to 2003, when many EE countries held referendums or signed treaties to join the EU. The timing offers a unique context in which there was no anticipation of the fall of the Eastern Bloc and, together with the data that offers unique rich information, identification is achieved. Approximately 30k academics from EE were identified, 3% of whom were migrants.

During this period, the Iron Curtain, a political boundary dividing Europe into two separate areas from the end of the Second World War in 1945 up until the end of the Cold War in 1991, was in place. As a result, it severely limited migration between the East and the West from 1950 up until its fall in 1991 (Van Mol and de Valk, 2016). The series of events that preceded border openings and the collapse of the Soviet Union led to the largest migration wave - in and from eastern Europe - ever since the events of refugee and forced migration of WWII (Bade, 2008). All in all, after the opening of the Iron Curtain in November 1989 marked by the fall of the Berlin Wall, immigration from eastern Europe started and surged in all categories, including migration of academics and scientists (Marshall, 2000). Thus, the collapse of the Iron Curtain induced new migration flows, and enabled and facilitated the migration of academics and researchers from Eastern Europe, the focus of this analysis. Note that the Eastern European authors identified in this period are tracked up until 2003, marking the year many Eastern European countries signed treaties or held referendums to join the EU and consequently the enlargement of the European Union in 2004. For example, after 2003, academic migration from eastern Europe to the United Kingdom increased through movements from the countries that gained access and membership to the European Union in 2004 (Burrell, 2010). In this research, we find evidence that academics were not able to anticipate the fall of the Berlin Wall and their possibility to migrate. As a result, there appears to be no manipulation of the network size and quality, throughout all types of migrants (between EE and out of EE), implying that migration did not induce networks, and thus enables us to reach the causal effect of networks on migration.

Using this context and these data, I test the assumption that the effect of the size and quality of pre 1989 academic networks, classified by location home, destination and foreign, on the probability to migrate, goes through two distinct channels: the cost and signalling channels, respectively. The cost channel is how the network characteristic reduces or increases the cost of migration,thus acting as a facilitator or a de-facilitator of migration. The signal channel on the other hand in which the network characteristic serves as a signal for the academic himself and his quality and his potential contribution and addition to the new host institution, thus also serving as a facilitator or a de-facilitator of migration.



Literature Overview

Network Theory and Social Capital Theory

- Jackson (2018); Bonacich (1991)
- degree centrality = size = community capital
- eigenvector centrality = quality = reputation capital

Iron Curtain and the Eastern Bloc

Effective barrier to migration (Van Mol and de Valk, 2016; Marshall, 2000)

Large migration flows post 1989 and a series of events (Bade, 2008; Burrell, 2010; Gozdziak, 2014)

Academic Migration and Selection into Migration

Becker et al (2020), De la Croix et al (2012), Borjas and Doran (2012), Bertoli and Ruyssen (2016), Gould and Moav (2014)

Contributions

- Adding to Blumenstock (2020)
- Testing the prevailing view (Carrington et al., 1996; Patel and Vella, 2013)
- Abstracting from the implicit assumption in the literature (McKenzie and Rapoport, 2010; Beine et al., 2011, 2015; Bertoli, 2010)
- Focusing on causal identification vs descriptive analysis done in CS papers using Big Data (Ricciarelli et al., 2020)
- The data! Mapping of networks and a wide range of controls and fixed effects

Figure showing the literature this research contributes to and draws from

SNEAK PEAK OF THE DATA!

The schema shows how everything is derived from the paper ID, information about the academic, his field, his co-authors (i.e. his network) and consequently information about his network (mainly their size and quality). Size is defined as the sum of co-authors, by location, from 1980-1988. Quality, on the other hand, is the average citation count and average rank of the co-authors by location (only for home average rank is used for definition issues), from 1980-1988. The main dependent variable is whether or not an academic migration post the fall and up to 2003, which is then classified into within-EE migration, out of EE migration, and no migration. The figure shows the motivation behind MAKG over other data sources. Even though the data is not perfect, especially when it comes to look at the quality of academics, MAKG is considered a better option, compared with other traditional scholarly data and other scholarly big data sources. Some descriptive results are important to note. Out of the approximately 30k academics from EE, 855 are migrants, 509 engaged in out of EE migration and 346 engaged in within EE migration. There is no evidence that academics strategically manipulated their networks in anticipation of migration due to the focus on Eastern European academics behind the Iron Curtain. Additionally, there was only within EE migration prior to the fall of the Berlin Wall giving further support for the Iron Curtain as a barrier to a migration from the East. The majority of academics' most frequent language of publication was English. Migrants tend to be older, have larger networks, smaller home network size, and larger foreign network size. The most famous destination is the United States, especially for mathematicians and scientists.

FINDINGS AND IMPLICATIONS

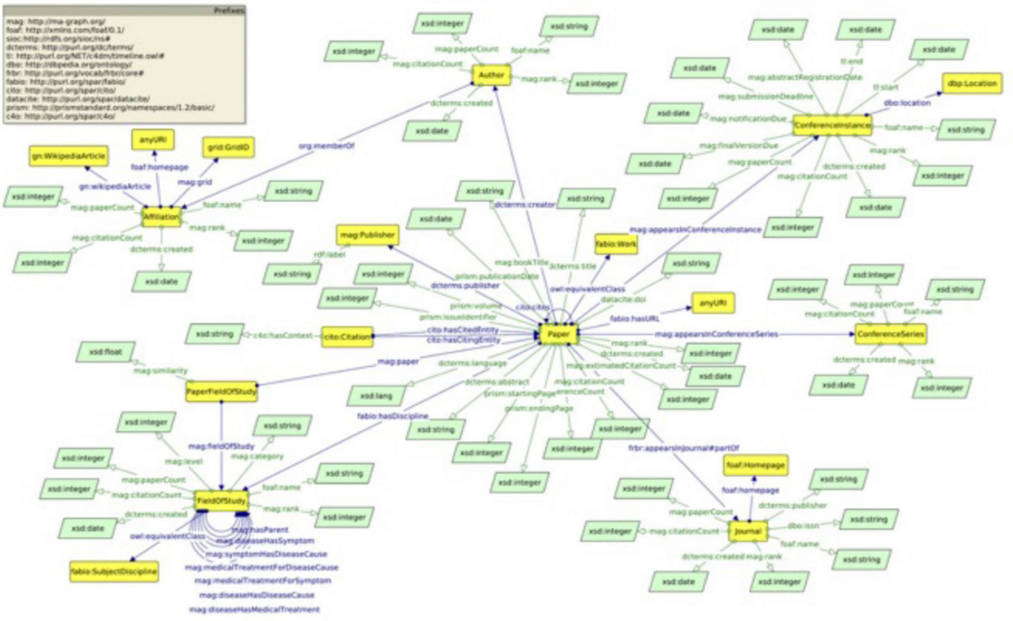
In this analysis, I find that an increase in the home network size (80-88) by one unit reduces the probability to migrate (1989-2003) by 0.1-0.05pp. Distinguishing between the types of migration, I find that an increase in home network size increases the chances of an academic not migrating compared to his chances of migrating to another EE. In fact, an increase in home network size by 1 unit, all variables constant, an academic is 1.034-1.071 times more likely to not migrate as compared to migrating to another EE , as the risk or odds are 3.4% -

7.1% higher. For the groups of those who migrated outside of EE vs those who migrated within EE, the evidence mainly implies lower chances of migrating out of EE compared to migrating within EE when home network size increases, aligning with the theoretical predictions. On the other hand, an increase in the destination network size for migrants increases the probability to migrate by approximately 7.7pp highlighting the lower costs of migration due to already established connections at destination and probably the easier the process of integration into the new host institution, aligning with the theoretical prediction. An increase foreign network size increases the probability of migration by 0.1pp, yet not statistically significant throughout all specifications. By distinguishing between the types of migration, evidence confirming the assumption that foreign connections are more likely to be close in terms of distance is found. This is because an increase in foreign network size by 1 unit increases the chances of an academic migrating within EE as compared to not migrating, as the risk or odds are lower by 1.5%. Similarly, an increase in foreign network size by 1 unit increases the chances of migrating within EE as compared to migrating out of EE as the risk or odds are 0.3%. For destination network size, as expected, the chances of not migrating versus migrating within EE is nearly zero.

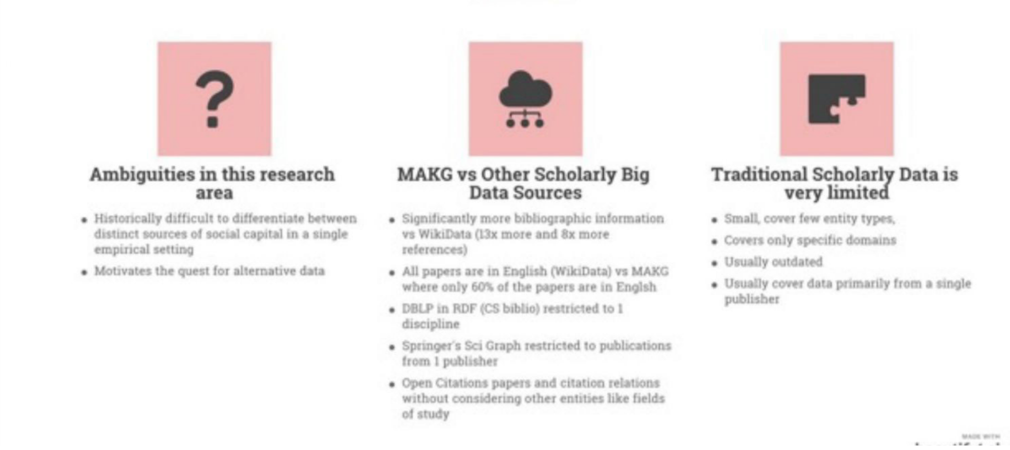
All of this confirms the fact that the cost channel mostly operates through network size in which greater networks at home increase the cost of migration, as leaving connections behind is costly, whilst establishing ones at destination reduces the cost of migration, as academic connections have already been established and would also ease integration in the host institution. For the foreign network size, the statistical insignificance and context does not help in disentangling which channel the effect operates through.

An increase in home network quality on the other hand, shows that the signalling channel marginally outweighs the cost channel as an increase in home network quality increases the probability to migrate. The effect is positive, statistically significant yet not economically significant (i.e. very small in magnitude), and thus the signalling channel outweighs slightly the cost channel, implying that an increase in home network quality acts as a signal for the academic's quality and, thus, a facilitator of migration more than a de-facilitator of migration, as "better" network at home is left behind. A decrease in home network quality increases the chances of not migrating versus migrating within EE, yet the effect is not very economically significant. Similarly, a decrease in home network quality decreases the chances of migrating outside of EE versus migrating within the EE. Looking at destination and foreign network quality I find evidence that supports the fact that better quality destination and foreign networks significantly increase the probability to migrate. However, the effect is economically not significant, being 0.0003pp and 0.0001pp. Considering these results, whilst having the results on destination and foreign network size at the back of our heads, they could imply that having a greater network at the destination or at a foreign country is more important that having a good connection at another country. This might be due to the fact that academics in EE were so segregated from the rest of the academic community worldwide that any additional connection would be of great help and would increase migration prospects, irrespective of the quality of that connection. Furthermore, an increase in foreign network quality increases the chances of an academic migrating within EE versus migrating out of EE as the risk or odds are 0.1% lower. This highlights that a greater foreign network quality, which is assumed to be usually in other EE countries, has an effect through the cost channel as leaving the region completely means loss of these foreign connections completely, thus, this acts more as a pull factor.

All of this confirms the fact that the signalling channel mostly operates through network quality, in which better networks by all locations act as a signal of the academic's quality, openness and options. However, the



Schema of Microsoft Academic Knowledge Graph showing how everything is derived from the papers themselves fact that many of the results are not economically significant and sometimes statistically insignificant highlights that size matters more than quality. This could be specific to this exact context, where academics were highly isolated from the rest of the academic world.



Ambiguities in this research area

- Historically difficult to differentiate between distinct sources of social capital in a single empirical setting
- Motivates the quest for alternative data

MAKG vs Other Scholarly Big Data Sources

- Significantly more bibliographic information vs WikiData (13x more and 8x more references)
- All papers are in English (WikiData) vs MAKG where only 60% of the papers are in Englsh
- DBLP in RDF (CS biblio) restricted to 1 discipline
- Springer's Sci Graph restricted to publications from 1 publisher
- Open Citations papers and citation relations without considering other entities like fields of study

Traditional Scholarly Data is very limited

- Small, cover few entity types,
- Covers only specific domains
- Usually outdated
- Usually cover data primarily from a single publisher

Motivating the use of Microsoft Academic Knowledge Graph as a data source for this research

As expected, prior migration does facilitate migration, especially if it occurred through an Eastern European country that became part of the EU in 2003. Looking at heterogeneous effects by broad disciplines offers some useful insights that are intuitive, novel, and confirm findings by previous, yet different, papers. There are no heterogenous home network size effects by broad discipline or field of study. In contrast, there are heterogenous effects of destination network size that are significant, statistically and economically, for Mathematicians, Computer Scientists and Engineers. This aligns with Borjas and Doran (2012) who argue that any Soviet Mathematician that tried to communicate with scholars outside of the Soviet Union, particularly in the US, could risk the potential attention from the KGB or even arrest. Thus, due to the extremely limited contact, an additional contact at destination would increase migration prospects from them, more than any other field, especially since they were of high quality/reputation. Additionally, an increase in foreign network size increases the probability of academics from the Arts and Humanities to migrate significantly more than academics from other fields. This could be explained by the fact that academics from fields that have larger network barriers and less quality signalling have a foreign network - which is a measure of openness, quality and options - that plays a more important role in facilitating migration. This aligns with the results from Becker et al (2021). Looking at the network quality, an increase in home network quality has a significantly different and positive effect on the migration probability of Mathematicians, Computer Scientists and Engineers. Evidence confirms that the signalling channel outweighs the cost channel. This happens more for these specific disciplines as they are the only ones with a significant interaction term when the home network belongs to the top 25%, whereas other disciplines would need their networks to be from top 10% so that the effect is significant. This hints at the reputation and quality of Mathematicians, aligning with Borjas and Doran (2012). There are no heterogeneous effects by destination network quality and only heterogeneous effects by foreign network quality for Arts and Humanities academics, aligning with the above interpretation for foreign network size.

CONCLUDING REMARKS

In conclusion, this research is important due to the various and vast contributions it offers to different strands of the literature. It first contributes to the current wave of research on human migration through the lens and perspective of big data (see Sirbu et al., 2020). However, it expands on this literature by focusing on a unique historical context that offers a much closer step to achieve the causal impact of networks on migration, thus it also expands on the literature focusing on migration post the fall of the Berlin Wall, the Iron Curtain and the dissolution of the Soviet Union. By focusing on academics, it contributes to the limited literature on academic migration and what shapes and affects their migration decisions (Teichler, 2015). It contributes to the vast and extensive literature on brain drain as the focus of this research is on academic migration. This research also provides a new empirical perspective on the determinants of academic migration paying particular attention to academic networks. As such, it also contributes to the strands of literature on the empirical relationship between networks and migration, which is an empirically hard task to do as mentioned before. The alignment between network theory and social capital theory also makes this research contribute to the empirical literature on social network theory.

This work is part of a micro-project done in collaboration between Paris School of Economics (Professor Hillel Rapoport and PhD student Donia Kamel), the University of Pisa (Dr. Laura Pollacci) and ISTI-CNR (Dr. Giulio Rossetti). This work is classified under the Migration Studies Exploratory under WP10, SoBigData++. It could also be classified under the Demography, Economy and Finance Exploratory.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

SoBigData Magazine is published under the

project N° 871042 | Programme: H2020 - INFRAIA

Duration: 01/01/2020 - 31/12/2023

Editorial Secretariat

info@sobigdata.eu

Editorial Board

Layout and Design

Beatrice Rapisarda

Copyright notice

Privacy statement

The personal data (names, email addresses...) and the other information entered in SoBigData Magazine will be treated according with the provision set out in Legislative Degree 196/2003 (known as Privacy Code) and subsequently integration and amendement.
Coordinator and Legal representative of the project: Fosca Giannotti | fosca.giannotti@isti.cnr.it

SOBIGDATA News is not for sale but is distributed for purposes of study and research and published online at http://www.sobigdata.eu/newsletter

To subscribe/unsubscribe, please visit http://www.sobigdata.eu/newsletter

SoBigData

SoBigData

www.sobigdata.eu