Agriculture and global food security
Economic and monetary union
Low-carbon economy and resource efficiency
Single market, growth, jobs and innovation
Public health, safety and security
Joint Research Centre
the European Commission's
in-house science service

Beyond official statistics to measure digital transformation

A big data approach to techno-economic segment analysis in the PREDICT project

INTERNATIONAL WORSHOP

MEASURES TO ENHANCE PRODUCTIVITY GROWTH
NEW DEVELOPMENTS

Universitat de València
Oct 30th 2017

Montserrat López-Cobo
Giuditta De Prato
Riccardo Righi
Sofia Samoili

EC JRC B6

Disclaimer:
The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Outline

Traditional PREDICT: Measuring the ICT sector and its R&D

Some highlights on PREDICT:

- ✓ What it is
- ✓ Why it's helpful

Recent extensions: Towards measuring digital transformation

Possible ways to measure ICT R&D & ICT content outside of the ICT sector

Delivered and work in progress:

- ❖ *Human capital: workers in ICT occupations*
- ❖ *IO framework*
- ❖ *Cross-border activity: international trade of ICT goods and services*
- ❖ **Techno-economic segments – TES**

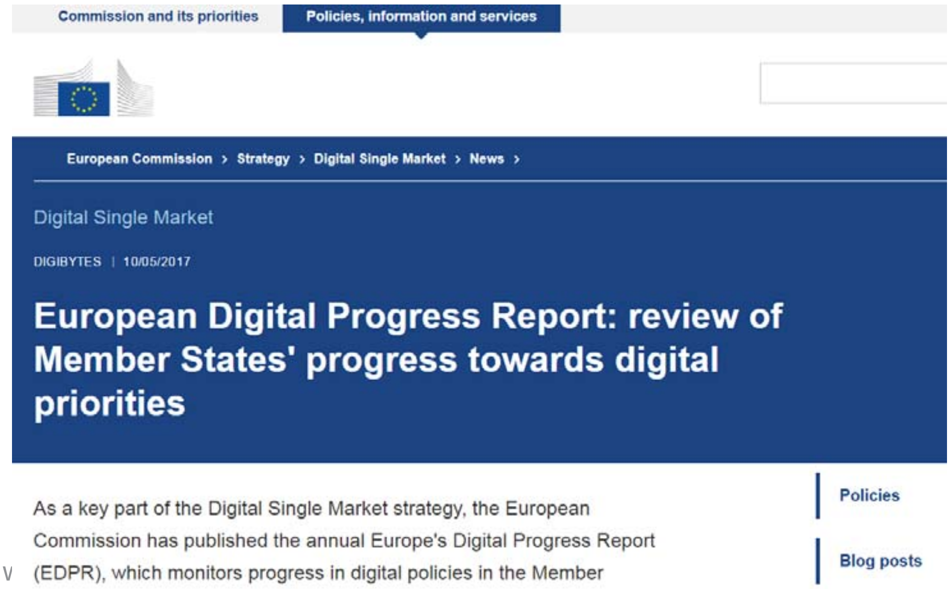
PREDICT 3: "Prospective Insights on ICT R&D‟ – 3rd phase

- Joint research project of the European Commission (EC) Joint Research Centre (JRC) and of DG CONNECT, follow-up of … PREDICT Arrangements since 2005!
- It produces comparable data on **ICT sector**, annual reports, exploratory analysis; it is based on latest available official statistics delivered by National Offices, Eurostat, OECD, etc.
- Designed to help policy makers understanding dynamics in the ICT sector and fostering its growth: PREDICT has become a unique source of information on the ICT sector and on ICT R&D in the EU and its global competitors.
- **2017 PREDICT Dataset**: the newest of the ever-improving PREDICT datasets, including the novelty of backwards reconstruction of the series from 1995, thus covering the period from 1995 to 2016.

PREDICT contributes to EDPR by DG CNECT:

- It used to provide data, original estimates and analysis to the **DAE Scoreboard** (2014, 2015, 2016..) for the evaluation of the DAE initiative
- Now providing data to the DESI index
- Providing data and a full chapter to the **European Digital Progress Report (EDPR)**

➔ **European Digital Progress Report published in May 2017**



Commission and its priorities
Policies, information and services
European Commission › Strategy › Digital Single Market › News ›
Policies
Blog posts

Digital Single Market

DIGIBYTES | 10/05/2017

European Digital Progress Report: review of Member States' progress towards digital priorities

As a key part of the Digital Single Market strategy, the European Commission has published the annual Europe's Digital Progress Report (EDPR), which monitors progress in digital policies in the Member

https://ec.europa.eu/digital-single-market/en/news/european-digital-progress-report-review-member-states-progress-towards-digital-priorities

Towards measuring the digital transformation: ICT across the economy

- Human capital: Workers in ICT occupations
- IO framework
- Cross-border activity: International trade of ICT goods and services
- Techno-economic segments – TES

ICT specialists

- **Definition**
  - **Conceptual** (based on OECD, 2004): workers who *have the ability to develop, operate and maintain ICT systems, and for whom ICT constitute the main part of their job*.
  - **Statistical** (Eurostat, 2015): based on ISCO occupations (ISCO-88 and ISCO-08)
  - **Source**: LFS, aged 16-74

ISCO-08	
133	ICT managers
2152, 2153	Electronics engineers, Telecommunications engineers
2166	Graphic and multimedia designers
2356	Information technology trainers
2434	ICT sales professionals
25	ICT professionals
251	Software and applications developers and analysts
252	Database and networks professionals
3114	Electronics engineering technicians
35	ICT technicians
351	ICT operations and user support technicians
352	Telecommunications and broadcasting technicians
742	Electronics and telecommunications installers and repairers

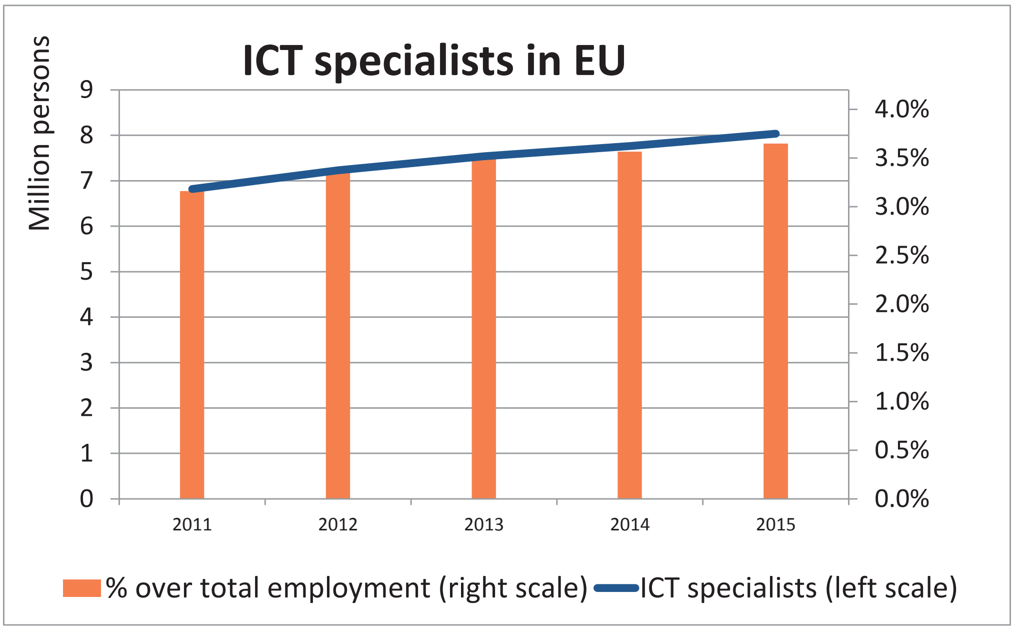
Source: PREDICT 2017: ICT Specialists

ICT specialists

- **Definition**
  - **Conceptual** (based on OECD, 2004): workers who *have the ability to develop, operate and maintain ICT systems, and for whom ICT constitute the main part of their job*.
  - **Statistical** (Eurostat, 2015): based on ISCO occupations (ISCO-88 and ISCO-08)
  - **Source**: LFS, aged 16-74

⇧ **Yearly number of ICT specialists**

⇧ **Share of ICT specialists in total Employment**



Source: PREDICT 2017: ICT Specialists

Measuring ICT Content across the economy

1. Investment (Gross Fixed Capital Formation)

- **ICT investment accumulation** as a factor of production: key input for economic growth.
- how much of the new value added in the economy is invested rather than consumed

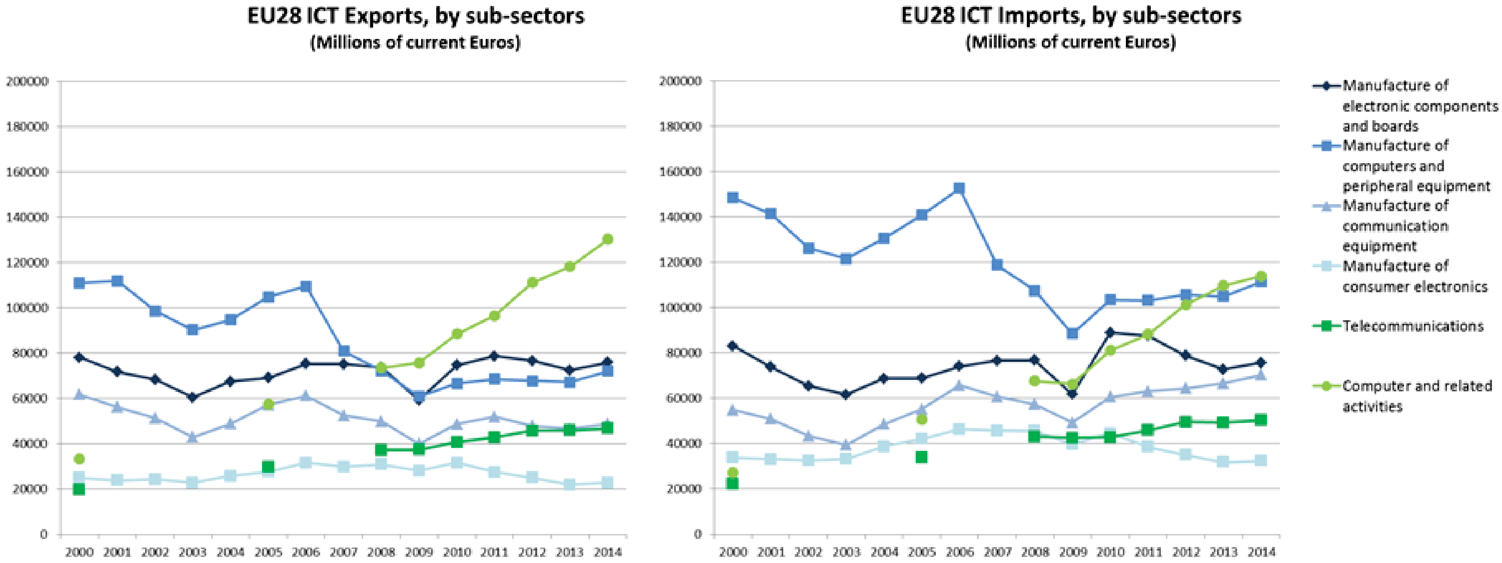
2. ICT content embedded in output (I-O framework)

- shows the role of ICT industry as seller and as buyer
- maps sectors according to the extent their output includes embedded ICT goods & services
- measure the use of ICT along the whole value chain by means of indirect and induced effects as well as in the final stage of production (direct effects)
- reflect both supply side and demand side effects

Digital Transformation

Intl Trade of ICT goods & services

Intl trade of ICT goods & services (2000-2014): imp/exp by subsect



EU ICT exports: increasingly services.

EU ICT imports: manuf. and services.

Data on Imports and Exports of ICT Goods and Services (2000 to 2014), for 41 relevant countries of the world, by sub-sector / end use category

Source: PREDICT 2017: ICT Trade

A holistic approach:
Network of international trade of ICT goods & services

Countries'(nodes) ***Degree***: the **number of connections** (trade partners).

- in-degree (D.IN): in-connections: trade partners for imports
- out-degree (D.OUT): out-connections trade partners for exports.

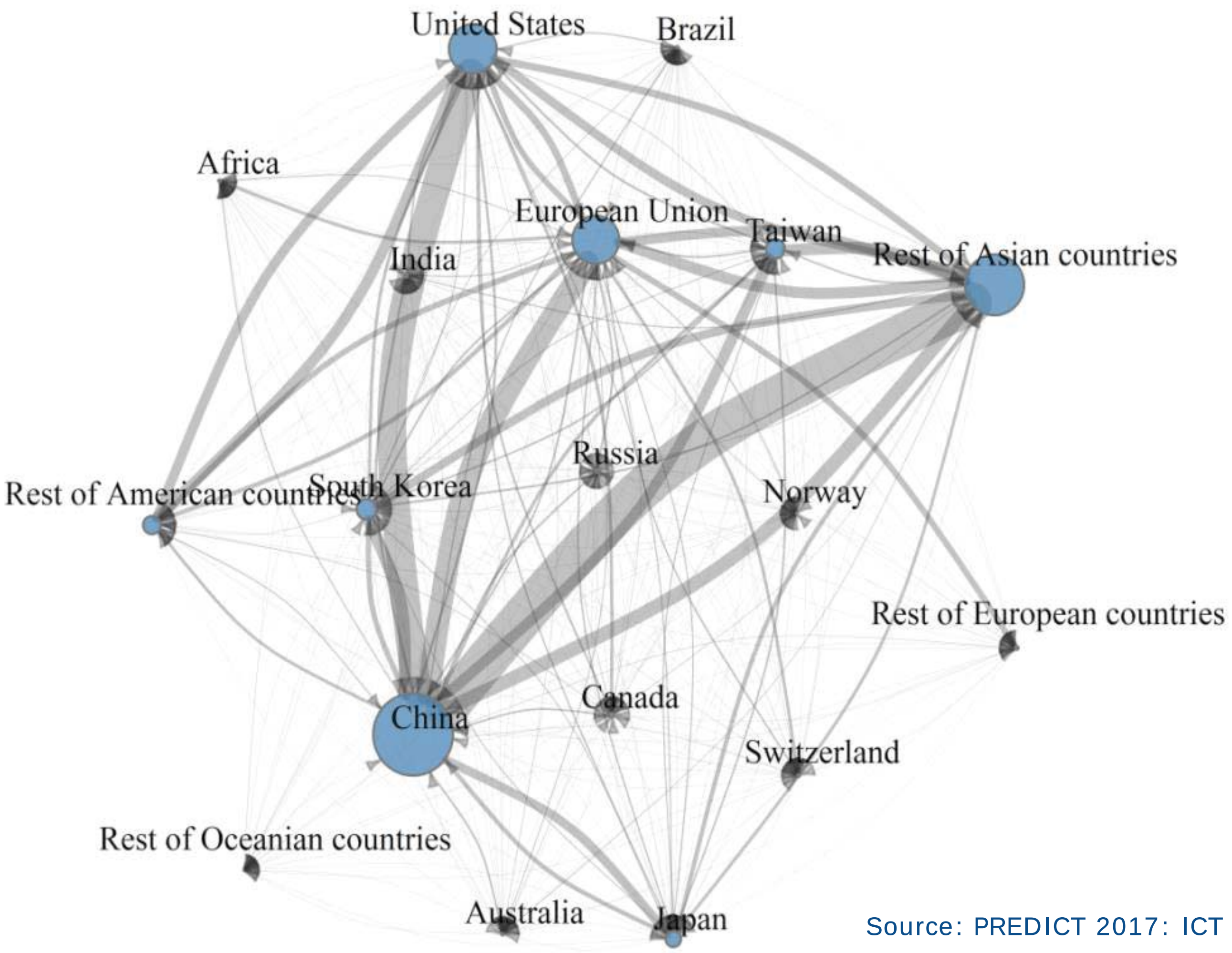
Countries' ***Strength***: the **sum of the weights** (total trade) of the connections of a node, considering the intensity of the relations that each node attains with all neighbours.

- in-strength (S.IN): total imports
- out-strength (S.OUT): total exports.

Countries' ***Weighted Betweenness Centrality*** (WBC): the **number of all shortest paths between any two nodes that pass through a given node**, thus considering topological properties, position and weight of each node, with respect to the entire network structure.

Multigraph: two countries may establish >1 connection per year: 7 groups of products/services

Intl Trade of ICT goods & services



Source: PREDICT 2017: ICT Trade

Digital Transformation

TES

European
Commission

Techno-economic segment analysis – TES

Measures to enhance productivity growth
NEW DEVELOPMENTS

Joint
Research
Centre

Fundación BBVA - IVIE
València, 30 October 2017

TES
An exploratory approach

✓ A new **target**: more than the ICT sector

- From sector to a **technological domain /ecosystem approach**
- Reflecting on **technological complexity/combinations**
- Exploring **internationalisation** of research, innovation, production and consumption

✓ Complementary **data sources** and **tools / techniques**

- text mining, semantic analysis, dynamic topic modelling, complex network analysis, community detection

⇨ **The objective is to analyse TES' size, characteristics and dynamics**

❖ Capture the **ecosystem(s)** – players, size, relations, dynamics, locations

❖ Describe the **global networks** and possibly the evolution in time

❖ Map the **hotspots** at EU or global level

❖ **Benchmark** or position – players, technologies, regions

❖ Capture the **technological dimension** & map the evolving technological map

❖ Spot **emergent TES** or **sub-domain** within a TES

European Commission

TES
Policy relevance

Why TES?

To provide DG CNECT with insights..

- ❑ for the **evaluation of specific H2020** lines
- ❑ directly connected to the **CNECT organigram**
- ❑ in line with the **PPP's objectives**
- ❑ about **measuring digitisation**, one of the key priorities of the Junker Commission: to develop the "digital sector" and to leverage the opportunities of digital technologies and services for the whole economy

Digital Transformation

TES
Policy relevance

European Commission

There seems to be many policy-relevant Techno-Economic Segments

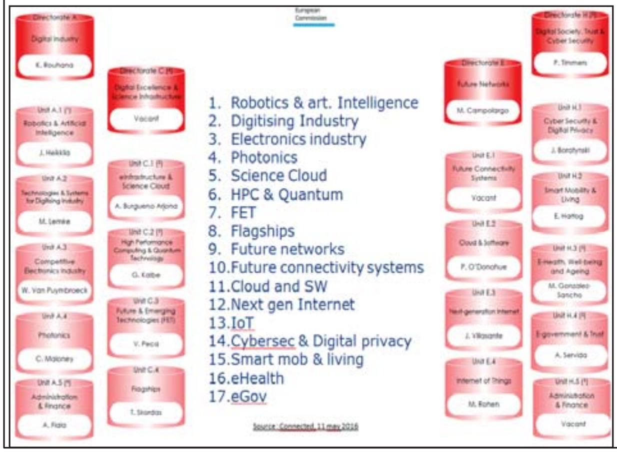
HORIZON 2020-led

Work Programme 2016 - 2017

Information and Communication Technologies

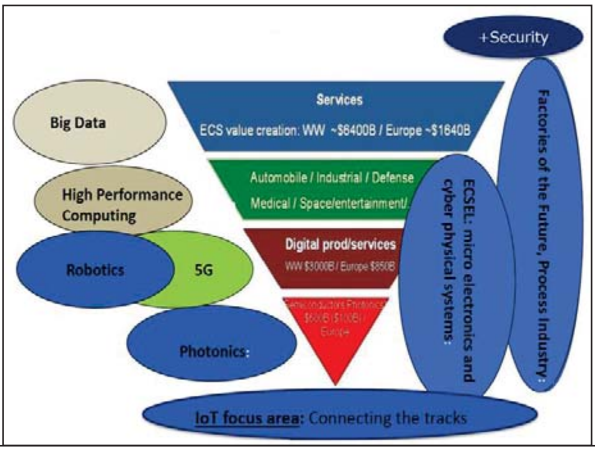
A new generation of components and systems
Advanced Computing and Cloud Computing
Future Internet
Content

Robotics and Autonomous Systems
ICT Key Enabling Technologies
- Photonics KET 2017
- Micro- and nano-electronics technologies



1. Robotics & art. Intelligence
2. Digitising Industry
3. Electronics industry
4. Photonics
5. Science Cloud
6. HPC & Quantum
7. FET
8. Flagships
9. Future networks
10. Future connectivity systems
11. Cloud and SW
12. Next gen Internet
13. IoT
14. Cybersec & Digital privacy
15. Smart mob & living
16. eHealth
17. eGov

Source: Connected, 11 may 2016


+Security
Big Data
High Performance Computing
Robotics
5G
Photonics:
Services
ECS value creation: WW ~$6400B / Europe ~$1640B
Automobile / Industrial / Defense
Medical / Space/entertainment/..
Digital prod/services
WW $3000B / Europe $850B
Semiconductors Photonics
$600B ($100B) /
Europe
ECSEL: micro electronics and cyber physical systems:
Factories of the Future, Process Industry:
IoT focus area: Connecting the tracks

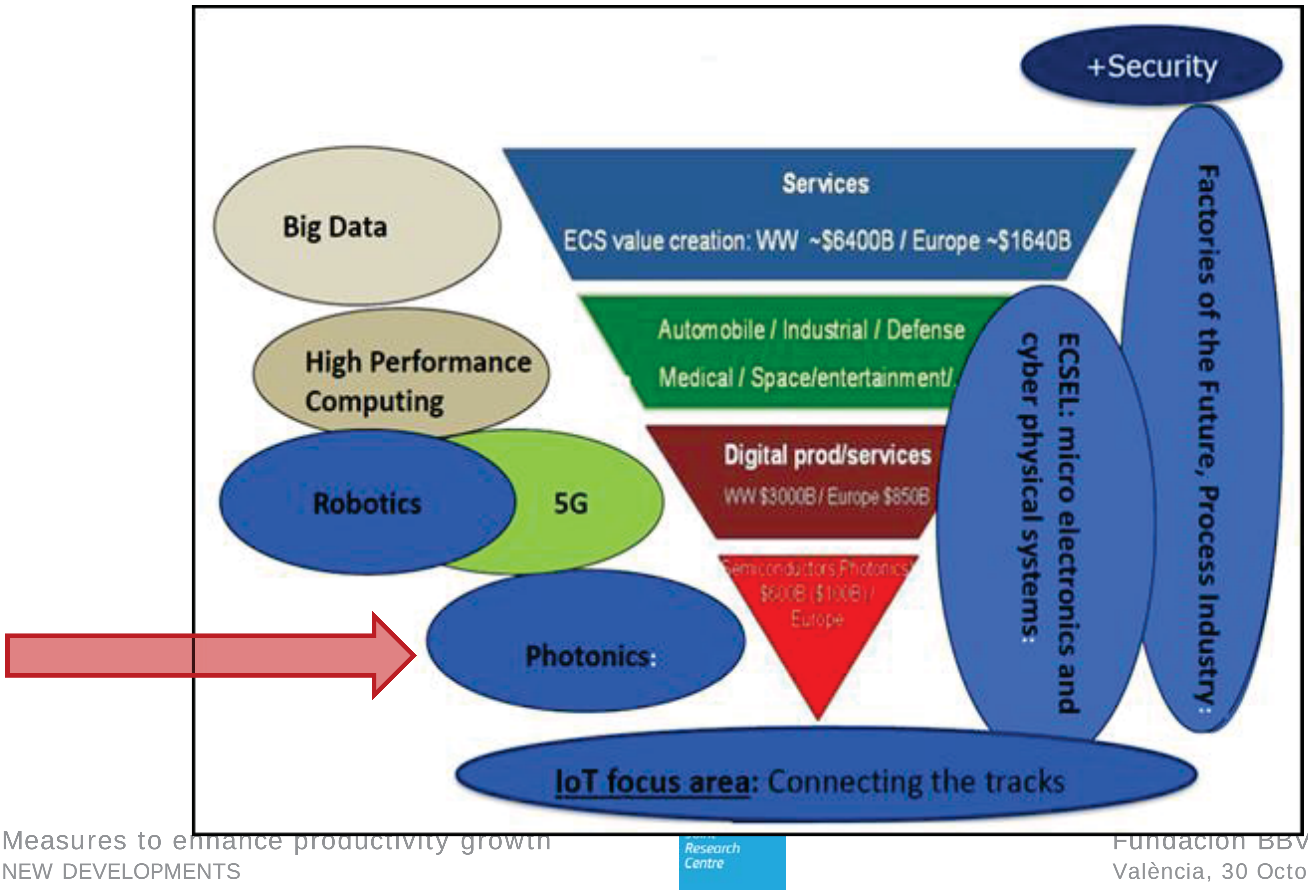
Or Techno-led: technological thesauri, industrial associations, standardisation bodies

Digital Transformation

TES
Selection of 1st segment

European Commission

After consultation with CNECT, the 1st proposed segment is PHOTONICS



+Security

Services

ECS value creation: WW ~$6400B / Europe ~$1640B

Automobile / Industrial / Defense

Medical / Space/entertainment/.

Digital prod/services

WW $3000B / Europe $850B

Semiconductors,Photonics/ $600B ($100B) / Europe

Big Data

High Performance Computing

Robotics

5G

Photonics:

ECSEL: micro electronics and cyber physical systems:

Factories of the Future, Process Industry:

IoT focus area: Connecting the tracks

From sector to segment

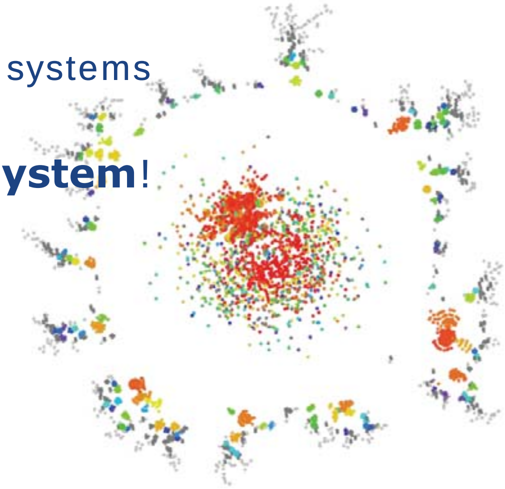
(TES) Techno-economic segment:

A grouping of companies, inventors, technologies, locations and stakeholders suitable to account for the whole ecosystem of a complex technology or otherwise labelled policy relevant "technology-based community"

- escaping industrial sector /subsector classification system
- escaping product classification system
- escaping intellectual property common classification systems

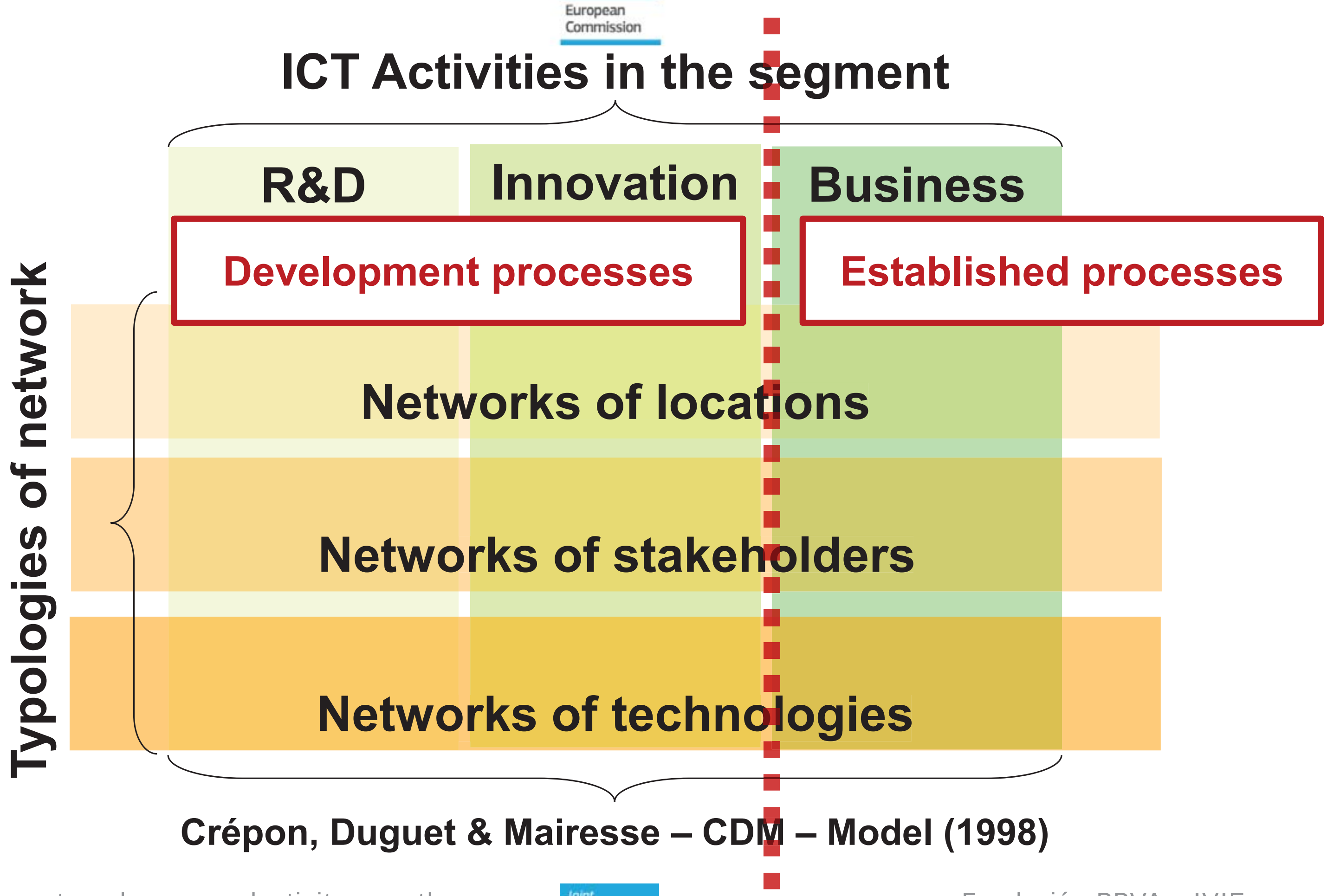
➔ an operational definition to target a **complex system**!

- described by its structure of interactions
- aimed at identifying evolving segments
- and detecting emergent behaviours/subdomains



Digital Transformation

TES
Simplified empirical framework



ICT Activities in the segment

R&D

Innovation

Business

Development processes

Established processes

Typologies of network

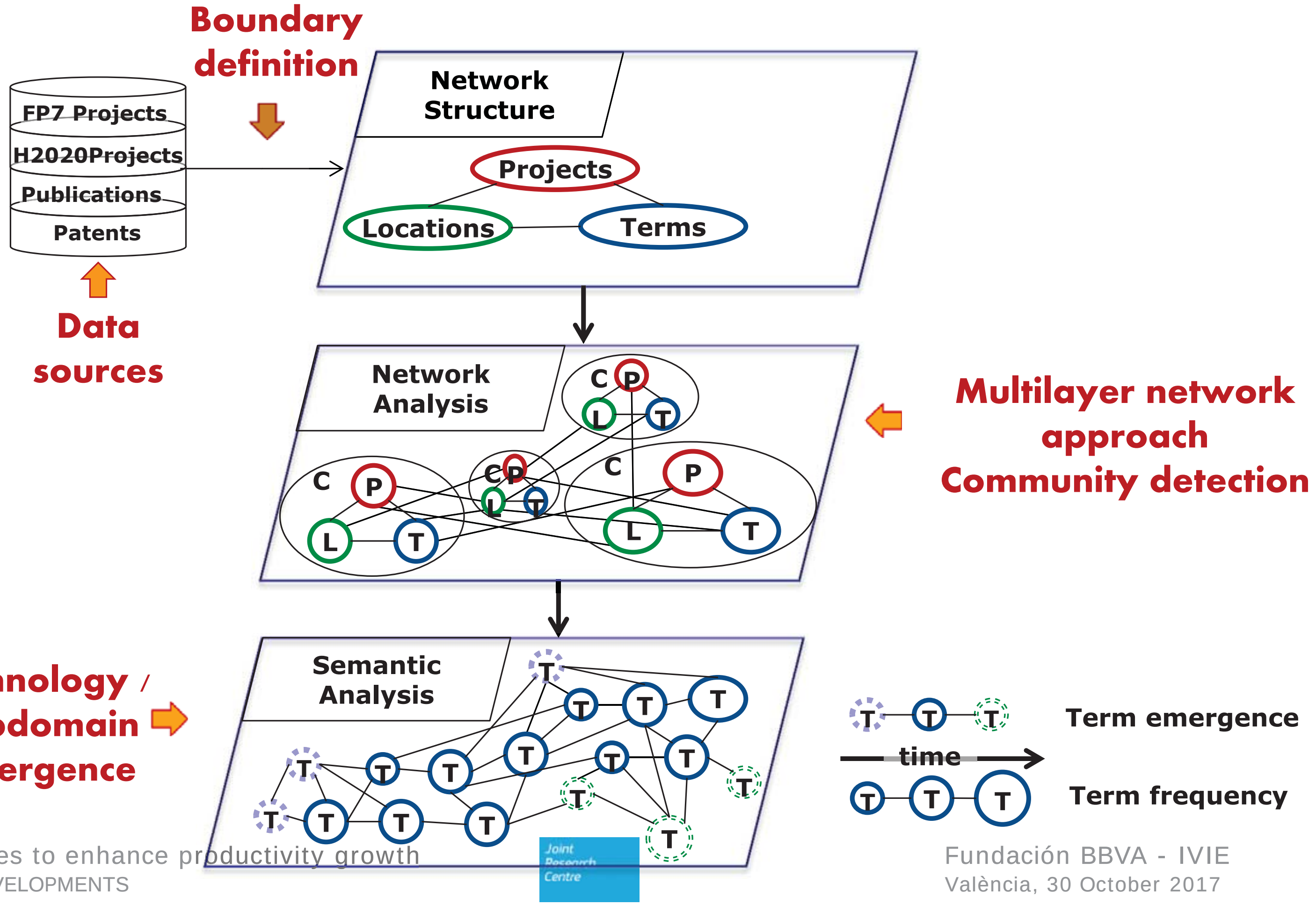
Networks of locations

Networks of stakeholders

Networks of technologies

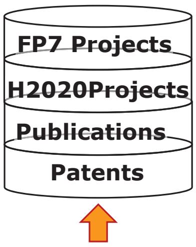
Crépon, Duguet & Mairesse – CDM – Model (1998)

TES
Framework of the approach



Digital Transformation

TES
Data sources


FP7 Projects

H2020Projects

Publications

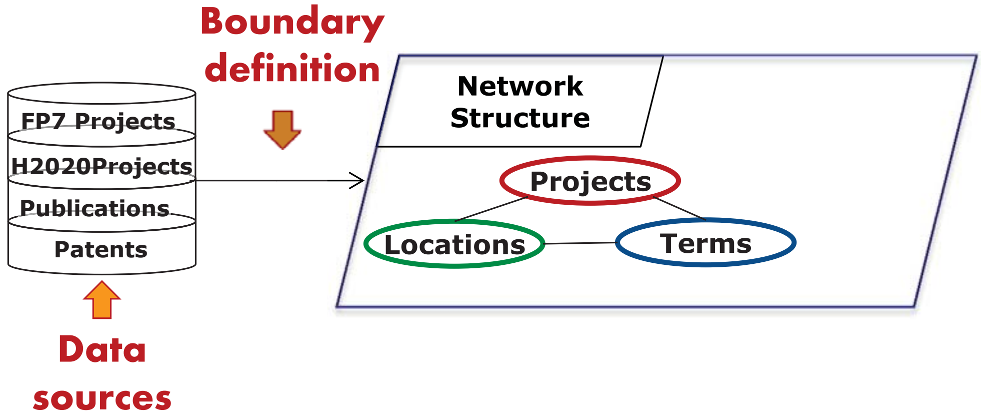
Patents

Data sources

- ✓ Allowing for a qualitative & quantitative approach
- ✓ Acknowledged data sources
- ✓ Covering the 3 activity dimensions: R&D, Innovation, Business
- ✓ "Mobilising" (scattered and heterogeneous) data

1. EU R&D funding: FP7 & H2020
2. Patenting behaviours: PATSTAT
3. EMM news sources (Europe Media Monitor)
4. Bibliometric production: Elsevier Scopus
5. R&D Centres location: Design Activity Tool by IHS iSuppli
6. Firm level data: ORBIS by Bureau Van Dijk
7. Venture Capital: VentureSource by Dow Jones
8. IHS resources via Goldfire semantic engine search tool
9. Unstructured data from market and industrial associations

TES
Framework of the approach



TES
Boundary definition

What defines the domain and boundaries of a techno-economic segment?

Top-down approach:
Are there thesauri, established with some consensus, about a techno-economic segment?

Bottom-up approach:
Is it possible to **reconstruct** the conceptual universe of a TES on the basis of some parts of its **production processes** (knowledge production, technical production, etc..)?

Example of a top-down approach:

US Photonics Buyers guide
62th Intl Ed
4k companies in over 1700 products

Company associations

PPP Photonic21

Example of a bottom-up approach:

Define a methodology
Baseline tool: Scopus by Elsevier
Photonics: 100K publications 2010-2016
Further refinements

TEXT MINING

I) Two step semantic approach to identify the relevant documents

1. **First query**: Scalable search on the DB of publications, FP7, H2020 projects and patents documents. **Query:** representative term **'photonic'** in abstract/description or title.
2. **Second query**: Scalable search on the DB to select documents that use same terms (most relevant 50 terms with highest value based on the Lucene scoring) → 21% of docs were added, comparing to the 1st query.

The **scoring function** is a combination of:

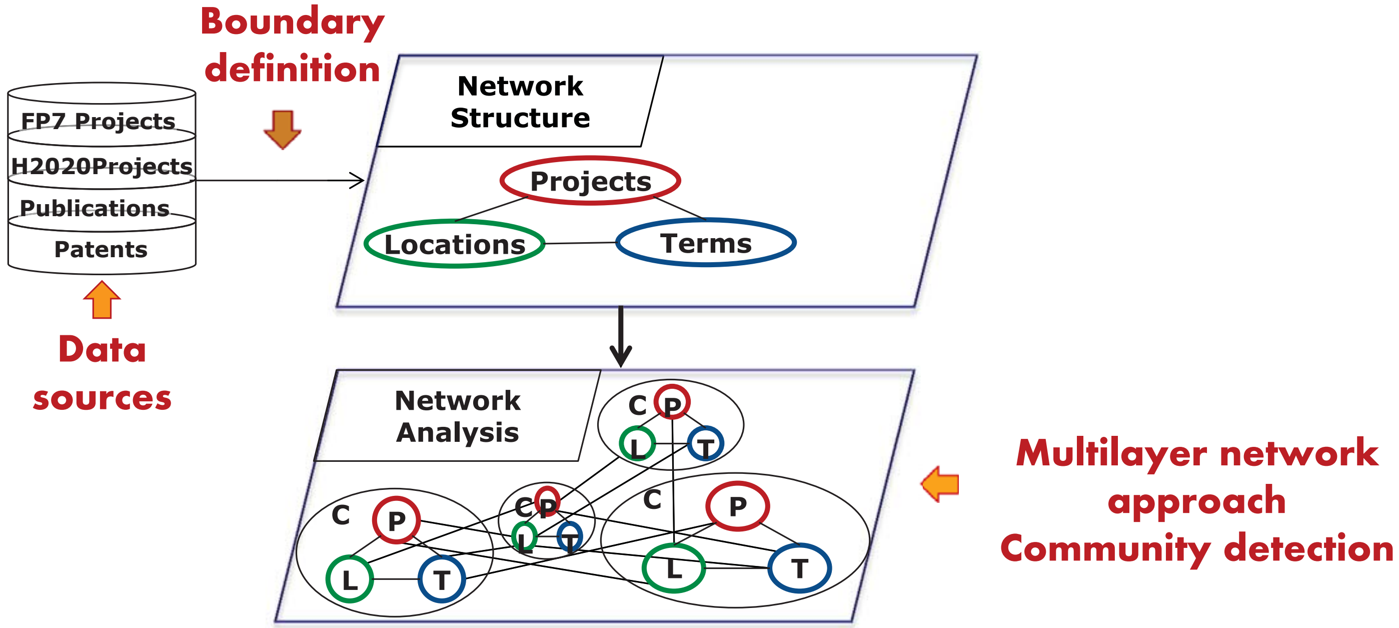
- the Boolean model,
- the *term frequency* and *inverse document frequency* indices (TF/IDF)
- field-length norm weight
- and the vector space model

II) Check terms against ad hoc Thesaurus with Elsevier TM

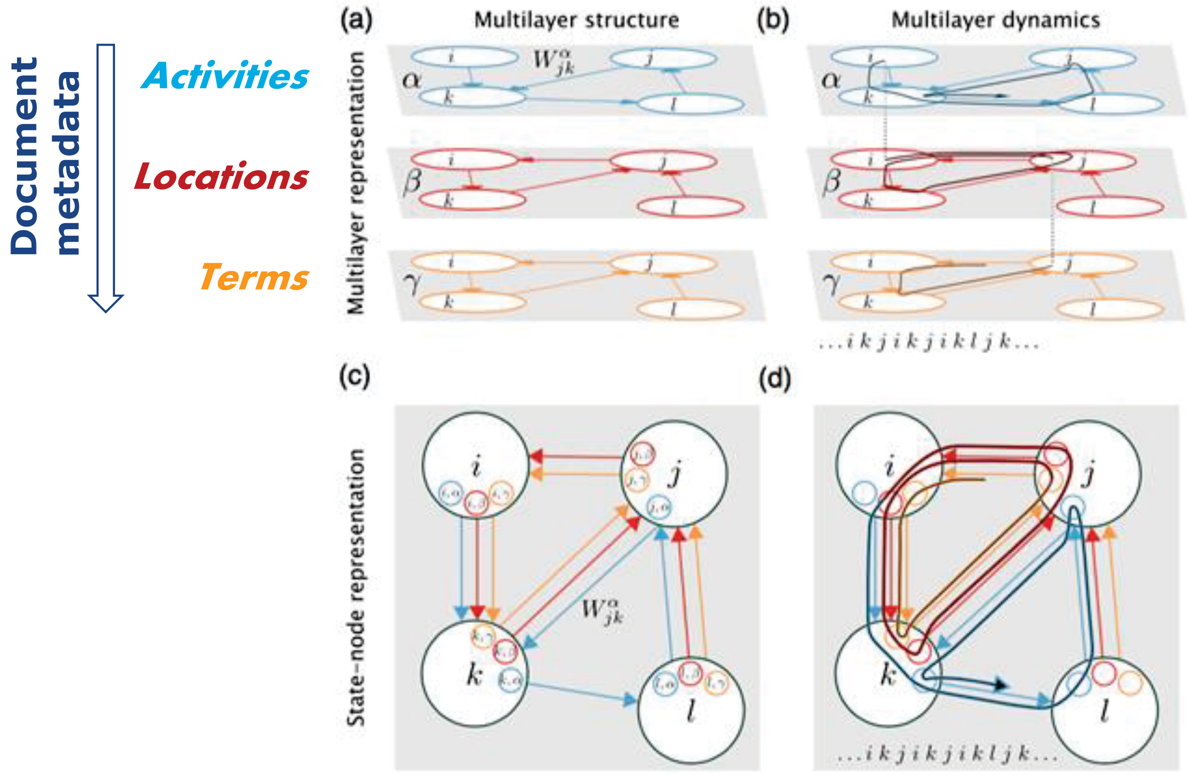
1. Identify **basic set of content relevant** for the domain of Photonics in Scopus
   - a) Potentially relevant technical terms: Noun phrases (NP), key phrases
   - b) Potentially relevant content: Compose basic set of documents
2. **Extract candidate phrases** for a photonics thesaurus from the basic set of relevant content
   - a) Extract and count NPs and use Inverse Document Frequencies (IDF) to downrate general terms
   - b) Extract and count document key phrases, apply IDF
   - c) Fine-tune term ratings by assigning different weights to NPs extracted from titles, abstracts and key phrases
3. **Select thesaurus candidate terms** in united set of candidate phrases
4. **Identify technical terms** in candidates (using technical / related thesauri)

Final Domain Vocabulary of 1,989 terms

TES
Framework of the approach



TES
Multilayer network approach



Document metadata

Activities

Locations

Terms

Source: De Domenico et al. (2015)

Three layers

- ***Players – Activities***: connected by co-participation in development processes
- ***Players – Locations***: connected by co-residence (local similarities/admin. proc.)
- ***Players – Terms***: connected by co-use of terms in their activities

Focus of the analysis

Identify overlapping communities of agents resulting from their interactions in different layers / dimensions

Tool: Community detection through the Infomap algorithm

(Fortunato and Hric (2016), Rosvall, Axelsson, Bergstrom (2009))

Why Infomap?

Infomap communities are formed by groups of nodes in which flow (of information, technologies…) is most likely to circulate.

This allows to shed light on how (and which) local dyadic interactions generate macro patterns of flows.

(Rosvall et al. 2009).

Value added of Infomap applied to multilayer complex networks (De Domenico et al., 2015):

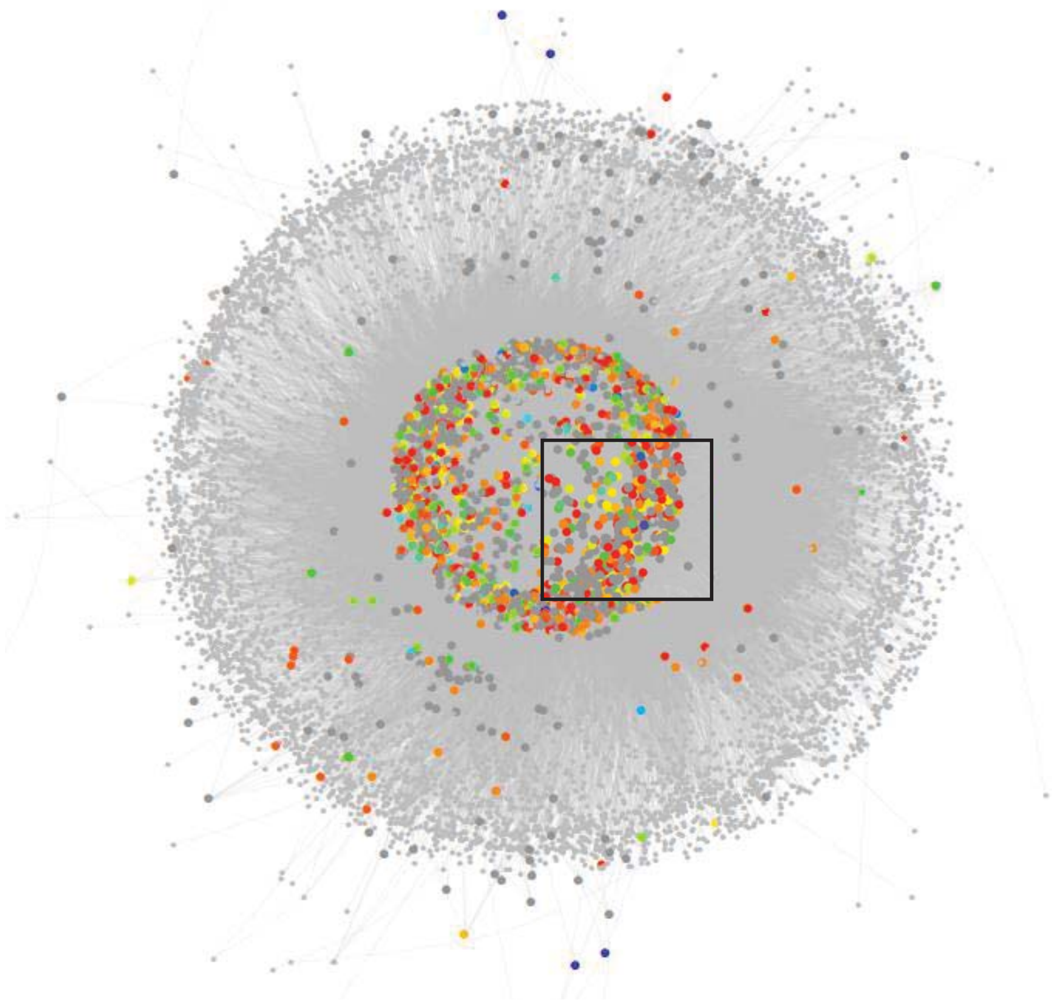
I. Analyses the **community structure** in each TES.

II. Analyses the **role of agents** involved with respect to the **whole network**, the **individual layers** and the **detected communities**.

III. Analyses the contribution of **each layer**, in the whole network and in each TES, to the generation of the total Infomap flow.

These aspects can be investigated in their **spatial dimension**.

TES
MLNA and Clustering



Light grey nodes:
events (activities / locations / terms)

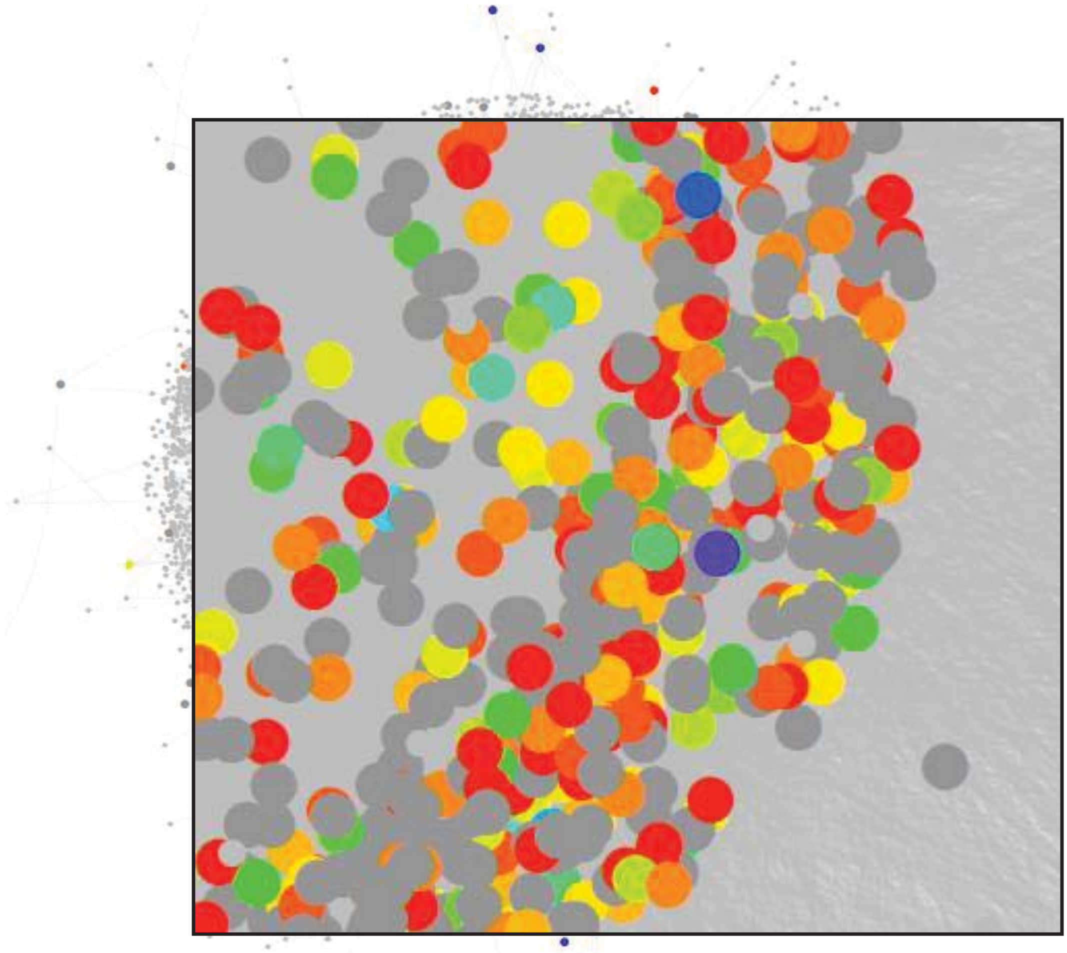
Dark grey nodes:
players belonging to the 1st (biggest) Infomap community -> the ‘generic cluster’

Other coloured circles:
nodes belonging to the 28 remaining (more specific) clusters.

Kamada-Kawai layout.

Digital Transformation

TES
MLNA and Clustering

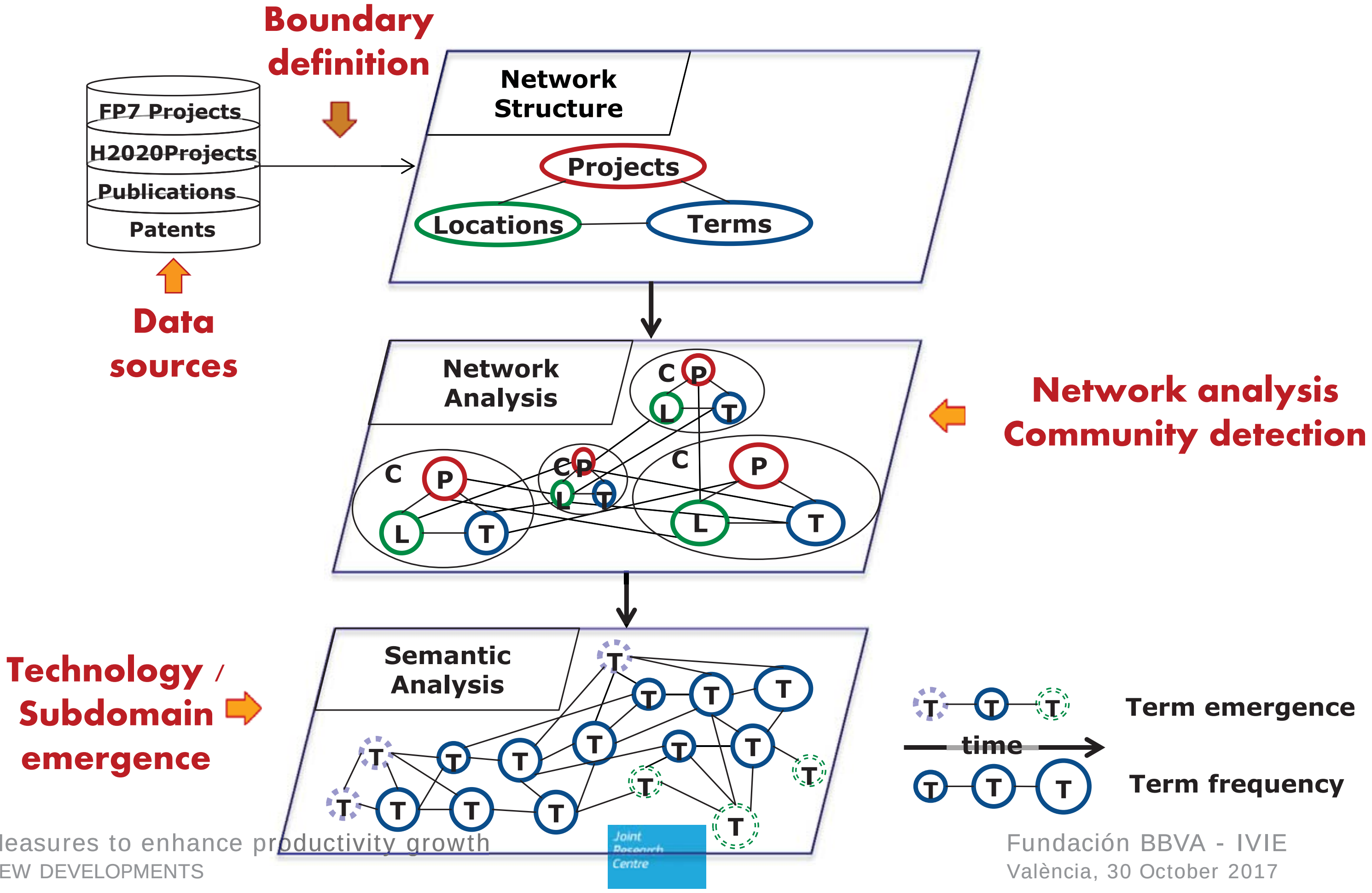


The **analysis** of each **cluster** provides information about:

- The type of **players** involved
- The amount and type of **information shared**
- The associated **events** they share (projects, common vocabulary/technologies used...)
- Their **spatial** characteristics

NEW DEVELOPMENTS

TES
Framework of the approach



TES
Semantic analysis

Objectives

- Identify segment subdomains/specializations.
- Analyze evolving & emerging topics, as proxies of current & emerging technologies.

How

- **Text mining**: to pre-process information from human-stored formats: remove common terms, punctuation, lemmatization, etc.
- **Natural language processing** methods: to detect part of speech, entity recognition, topic modelling.
- **Statistical tools**: dimensionality reduction -> derive the most significant set of terms occurred from each community for the current/emerging technological trends.
- **Probabilistic models**: to describe and assess **current**, **evolving** and potential **emerging** topics: Dynamic topic modelling (DTM), n-gram Markov Chain Model (MCM).

¡Gracias!

montserrat.lopez-cobo@ec.europa.eu

Visit out site

https://ec.europa.eu/jrc/en/predict