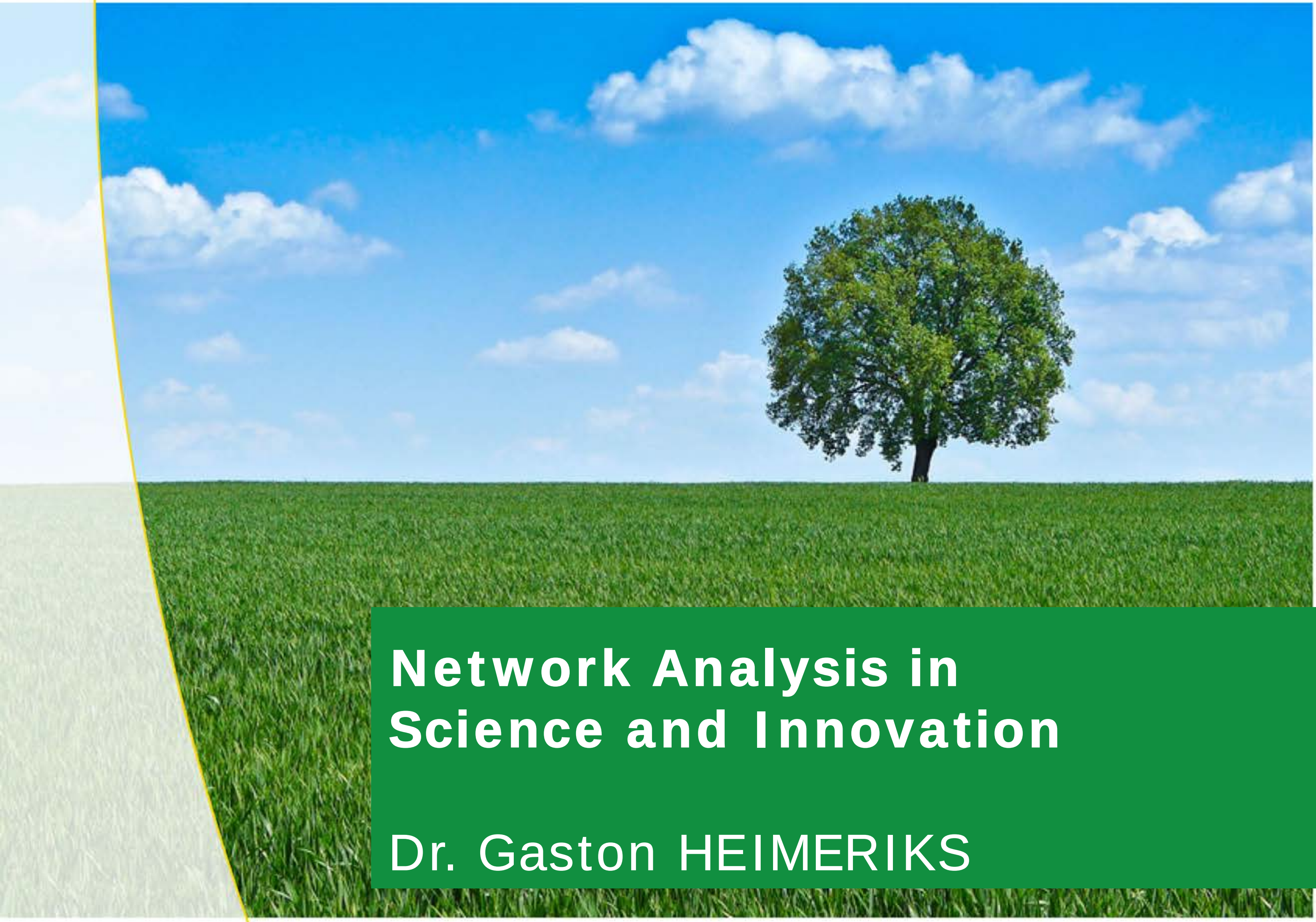
Universiteit Utrecht

Copernicus Institute of
Sustainable Development



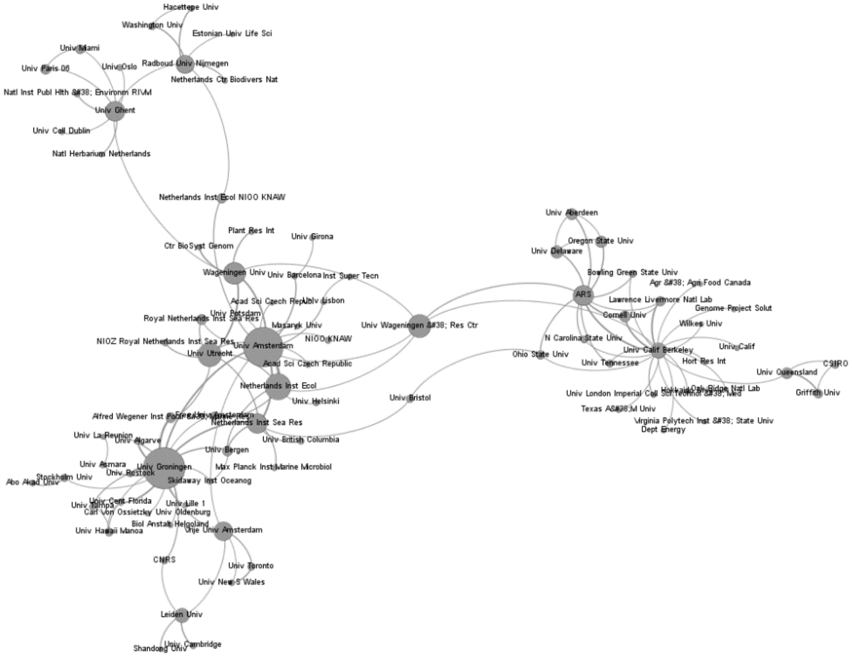
Network Analysis in Science and Innovation

Dr. Gaston HEIMERIKS

Outline

- Why networks?
- A recent application in innovation studies
- Scientometric networks
  - Data
  - Tools
- Data formats for network analyses

Why networks?



Different ways to look at reality

Attributes

- All Possess One or More Properties as an Aggregate of Individual Actors
- Examples: Firms, Men, Developed Countries

Networks

- Set of Connected Units: People, Organizations, Networks
- Examples: Friendship, Organizational, Inter-Organizational, World-System, Internet

Social Network Analysis

We measure Social Network in terms of:

1. Degree Centrality:

   The number of direct connections a node has. What really matters is where those connections lead to and how they connect the otherwise unconnected.

2. Betweenness Centrality:

   A node with high betweenness has great influence over what flows in the network indicating important links and single point of failure.

3. Closeness Centrality:

   The measure of closeness of a node which are close to everyone else.

   The pattern of the direct and indirect ties allows the nodes any other node in the network more quickly than anyone else. They have the shortest paths to all others.

Social Network Analysis (II)

4. Path length

The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.

5. Structural equivalence

The extent to which nodes have a common set of linkages to other nodes in the system. **Positional!**

6. Structural hole

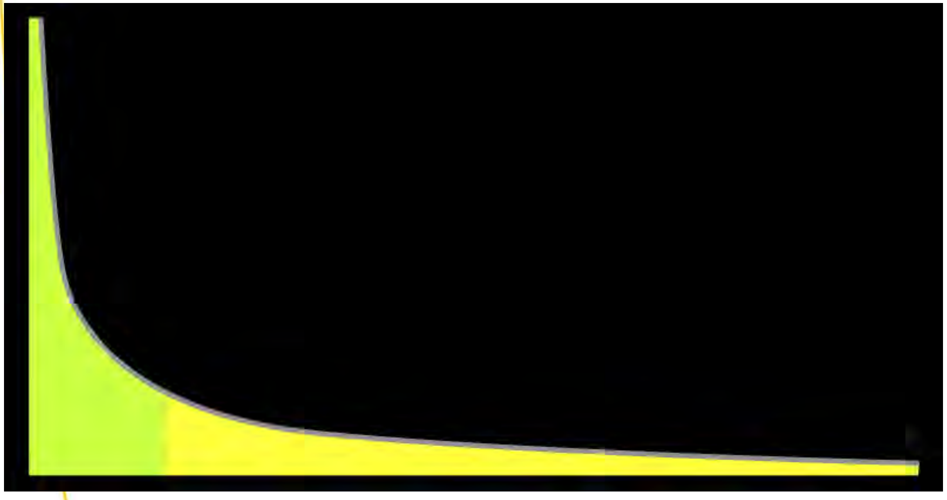
Static holes that can be strategically filled by connecting one or more links to link together other points (“niche”)

Network Analysis is in the context of science and innovation studies

- Network Analysis is in the context of science and innovation studies the application of **powerful statistical tools** and **analytical techniques** to uncover the **structure** or **development** of science based on the **relations** between **specific entities or units**.

- It can be applied to all units associated with science like
  - publications, disciplines, journals, institutions and researchers….
- Most likely, the results are plotted in a two- or three dimensional representation (a map). Therefore it is often referred to as ‘Mapping Of Science’.

Skewed distributions



social phenomena exhibit powers rather than linear progressions, and thus in most cases a power law is not so much surprising as it is overwhelmingly expected.

For example, the distribution of cited references in each field shows this power law distribution

- Can we explain these distributions by looking at the attributes or by looking at the network properties?

preferential attachment

A **preferential attachment process** is any process in which some quantity, typically some form of wealth or credit, is distributed among a number of individuals or objects according to how much they already have, so that those who are already wealthy receive more than those who are not.

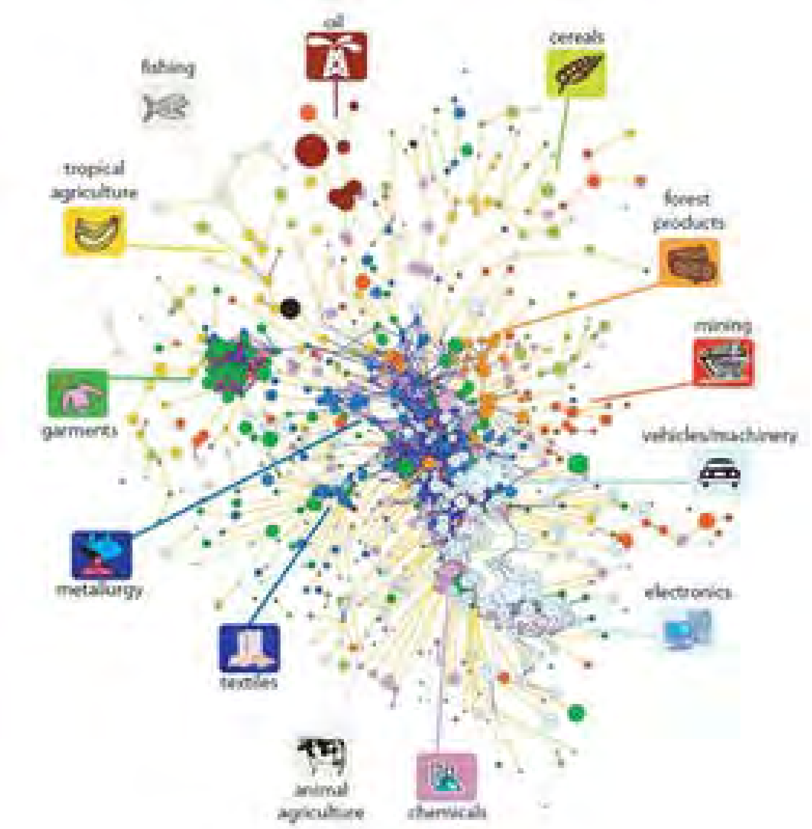
"Preferential attachment" is only the most recent of many names that have been given to such processes. They are also referred to under the names "cumulative advantage", "the rich get richer", and, less correctly, the "Matthew effect".

Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.

We will never understand complex (innovation) systems unless we map out and understand the networks behind them.

A recent example

Relatedness: Product space



A network of the relatedness between products, providing insight into the economic question of why some countries can quickly climb the manufacturing ladder, while others fail to develop more sophisticated products.

Product space

- In a recent issue of ***Science***, physicists César Hidalgo and Laszlo Barabási from Notre Dame, along with economists Bailey Klinger and Ricardo Hausmann from Harvard University, have presented a network of what they coined as the "product space."

- In the network, connections show the distances (similarities) of the exports a pair of products of a country. Their results show how the types of products a nation produces and exports determines the probability of that nation developing more competitive products, thus influencing its overall economic wealth and growth.

To make a product you need more than Capital and Labor

Public Inputs

Tradable Inputs

Certifying Body

Leather

Trade Agreements

Roads

Technical Education

Tanner

Ports

Power

Leather Cutters

Labor Skills

Tax Regulation

Leather Pressers

Sawing

Norms

trust

Sole Making

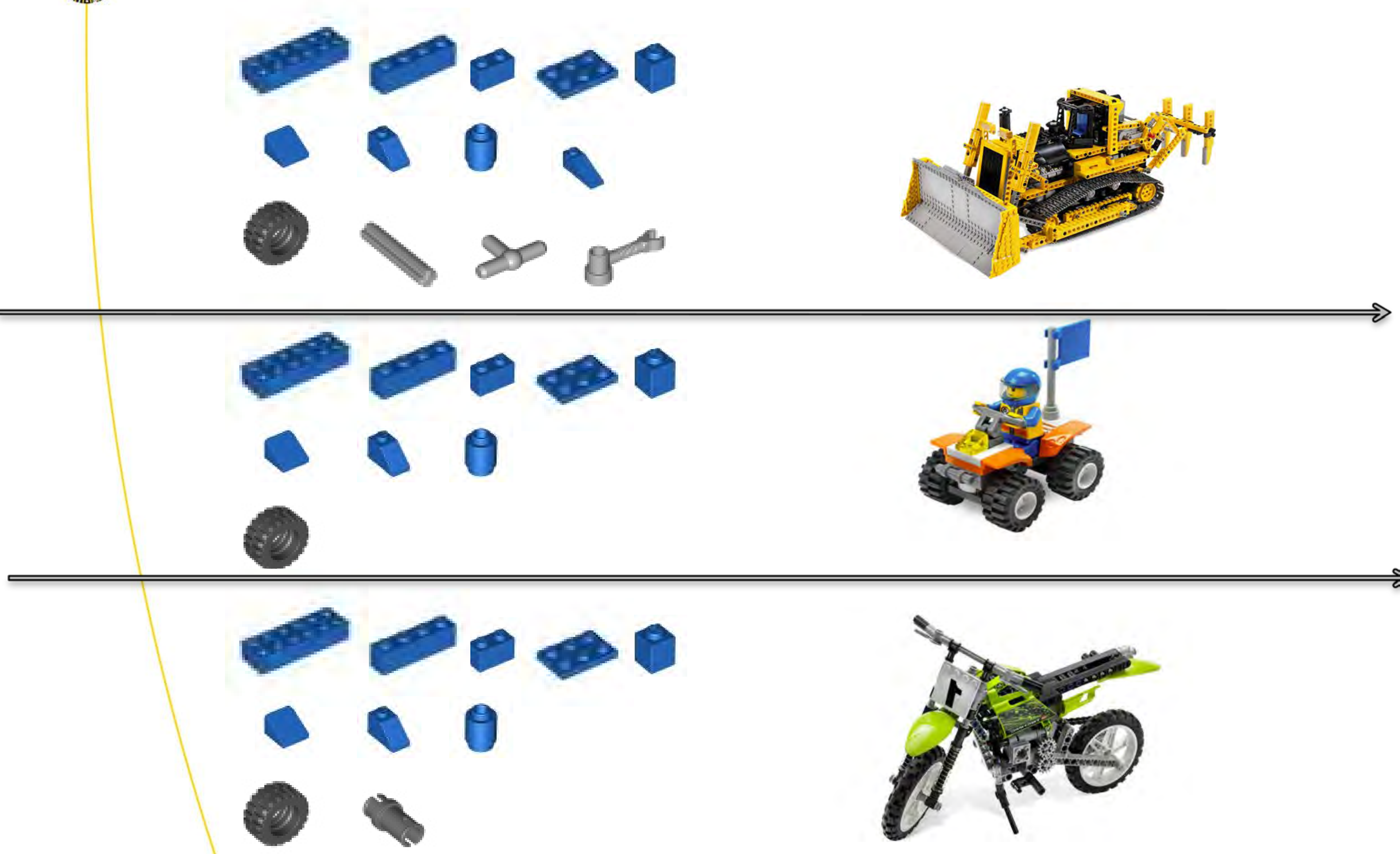
teamwork

Manufacturing & Management Certifica

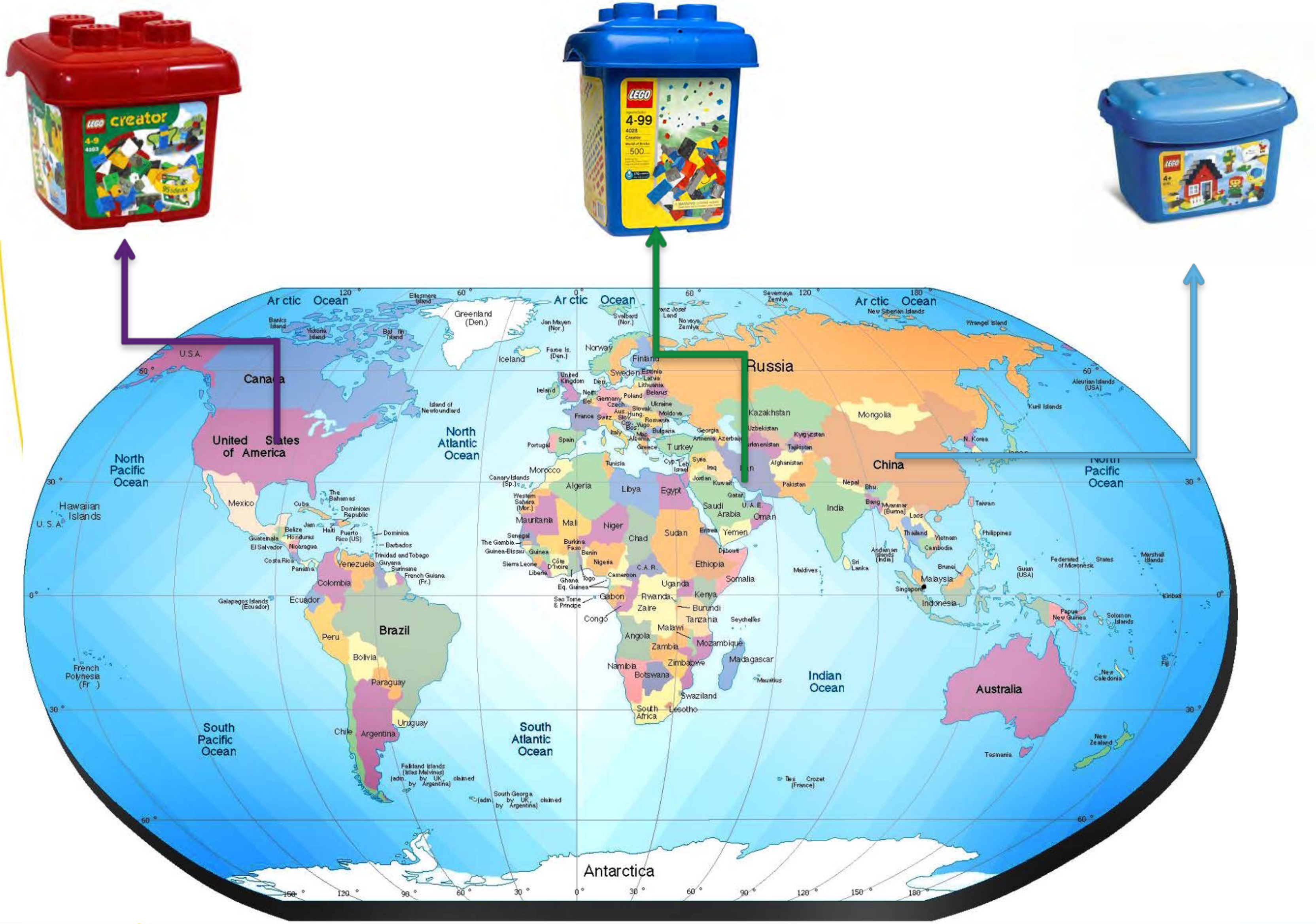
Shapers

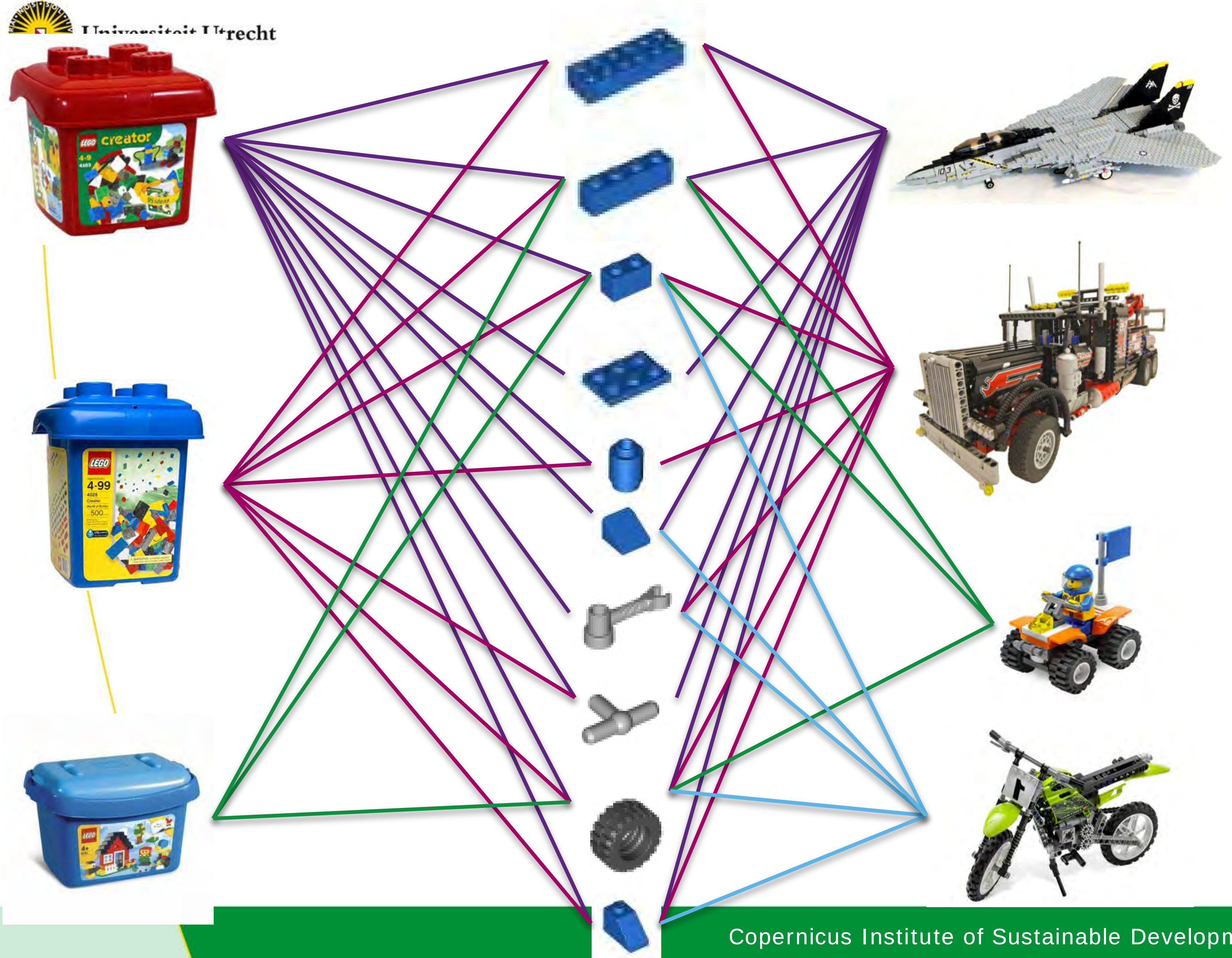
Private Inputs

Capabilities and Products



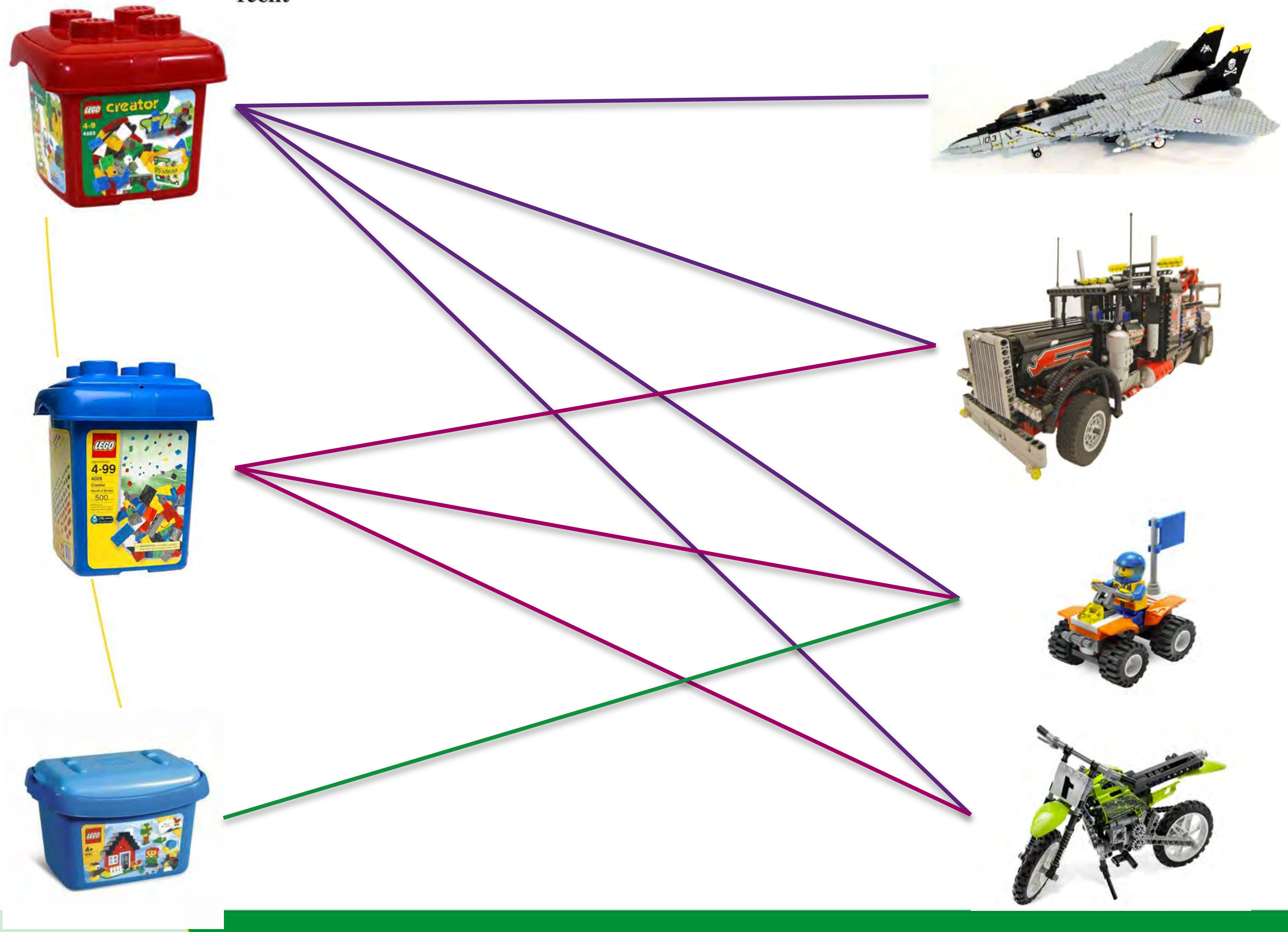
COUNTRIES and capabilities




Countries
Capabilities
Products

Countries

Products



Here I will show that within this view it is possible to:

MEASURE	-***measure*** the evolution of technological complexity
UNDERSTAND	-Demonstrate that ***economic complexity is a fundamental determinant of growth, innovation and the ability to address societal challenges***.
MODEL	-Model ***how a country's economy develops*** and show that this evolution is compatible only with a disaggregate view of the world.
DERIVE POLICY	-Derive **policy implications** from the view in context of existing measures

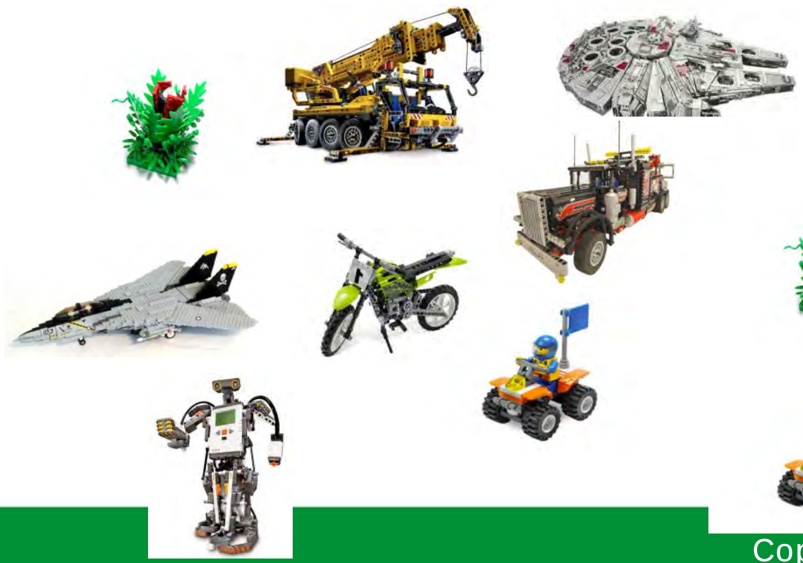
1

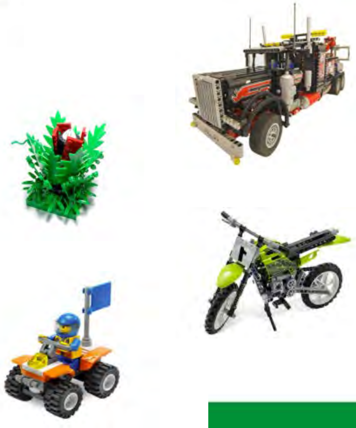
Measure

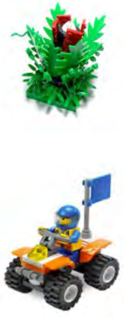
1.- A Country

2.- with a great diversity of Legos (capabilities)

3.- Can make many products



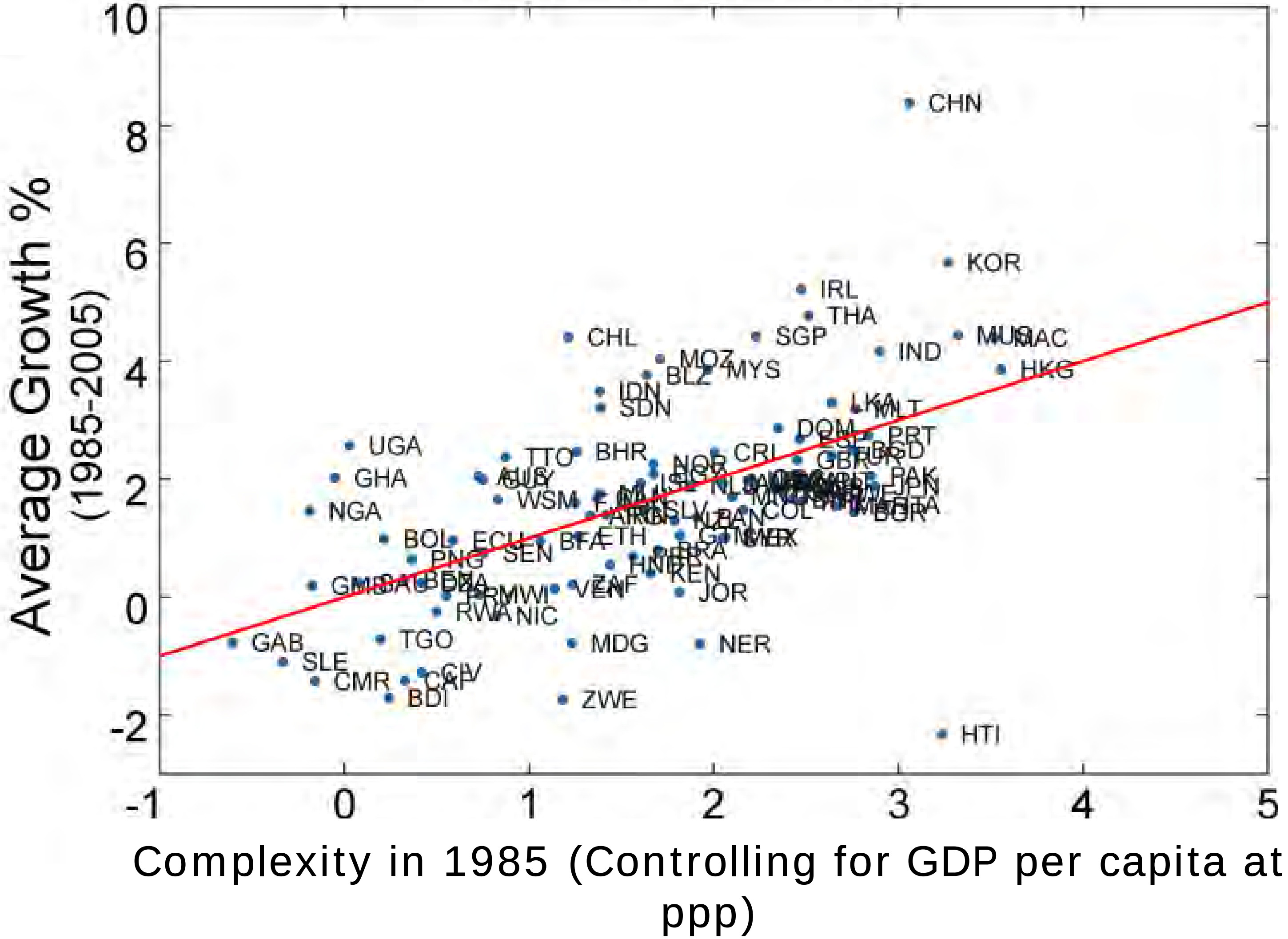




Universiteit Utrecht

2 understand

COUNTRIES APPROACH A LEVEL OF INCOME WHICH IS DETERMINED BY THE COMPLEXITY OF THEIR CAPABILITIES!!!

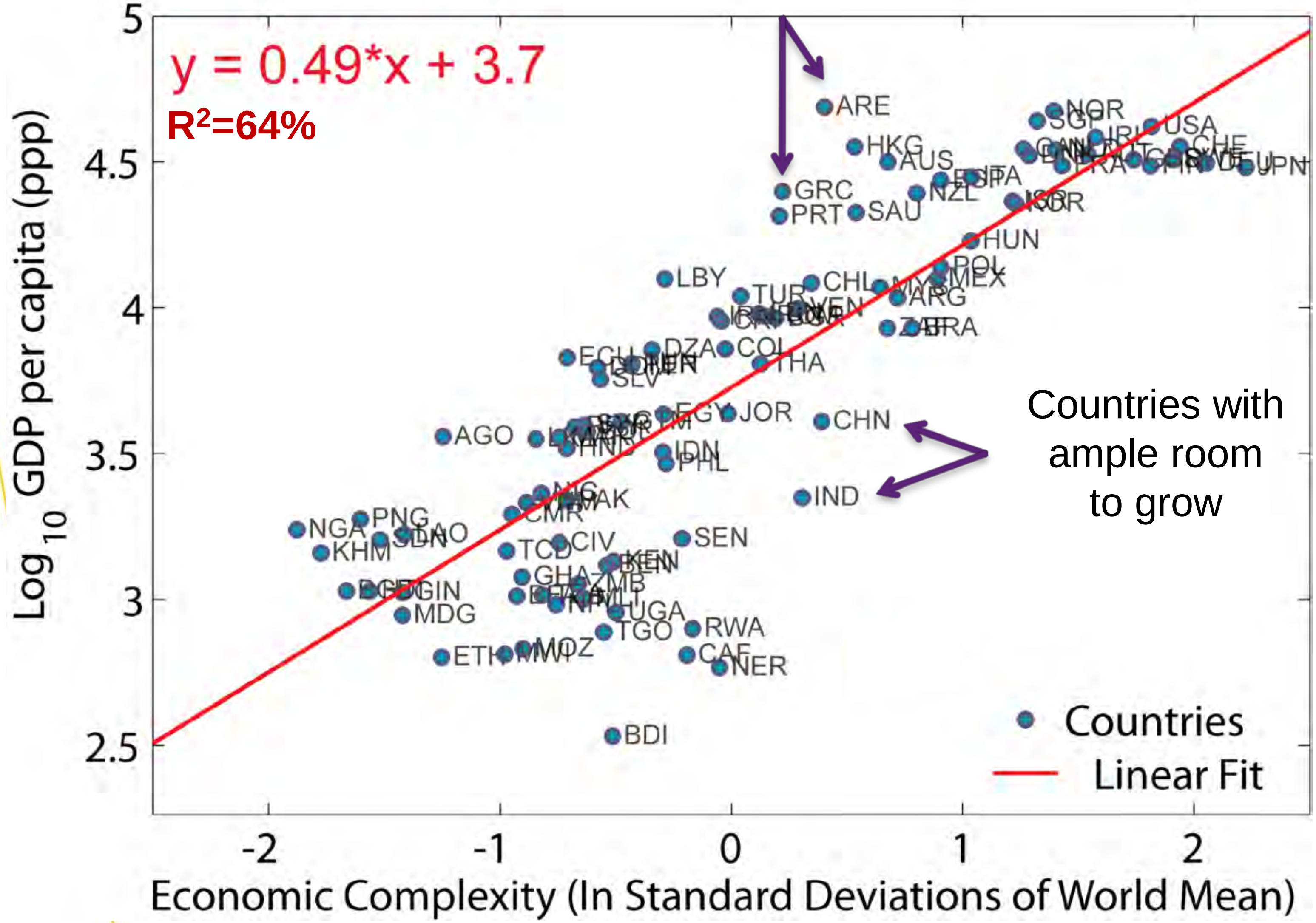


Hidalgo, Hausmann (2009) *PNAS* 106(**26**):10570-10575

Countries before the Financial Crisis

Universiteit Utrecht

Countries with an Income that “cannot” be sustained by the complexity of their economies

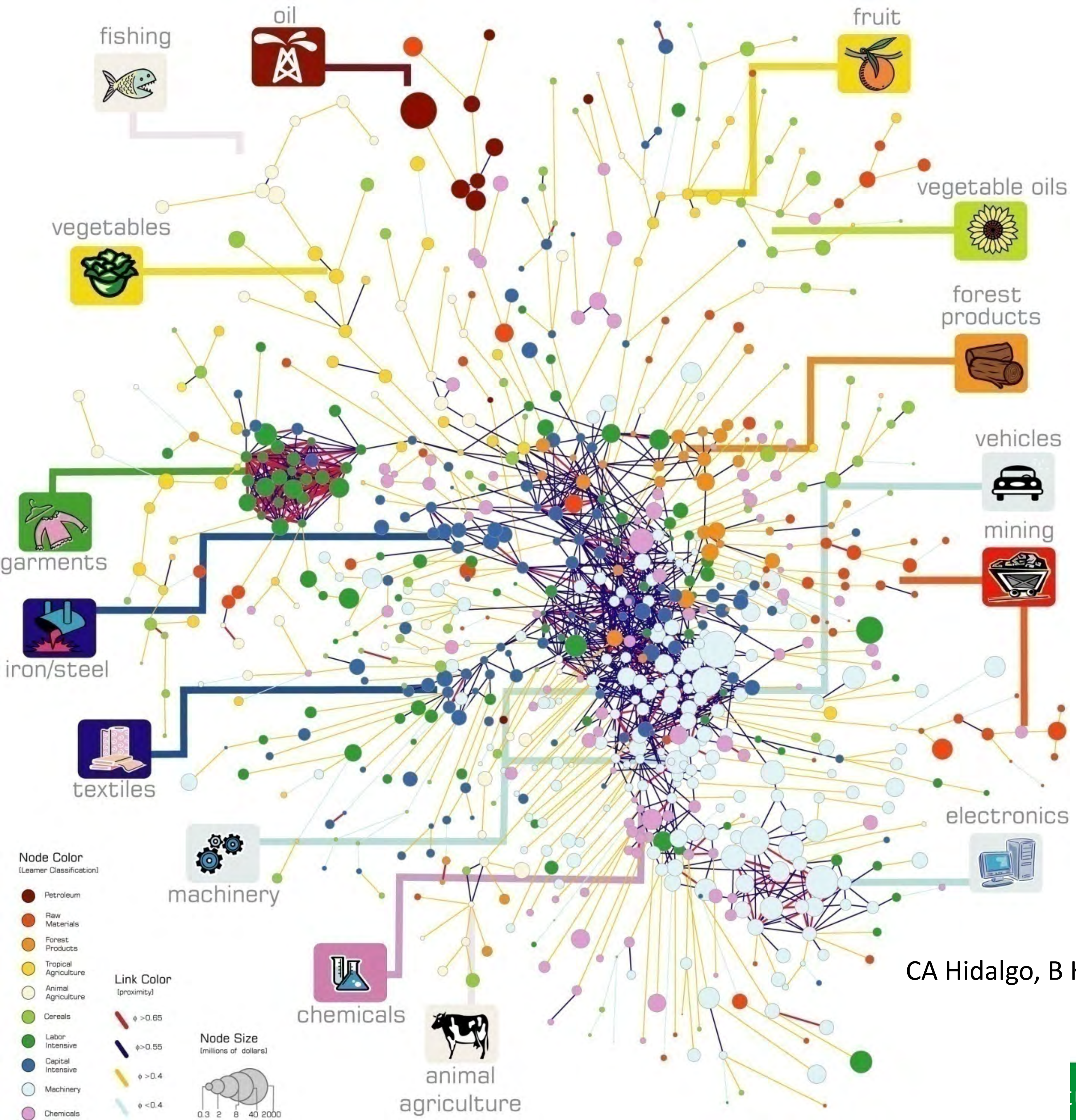


Countries with ample room to grow

Data for 2005, Source Bacii Dataset from Cepii. Products disaggregated in HS-6 (5109 product categories)

3 model

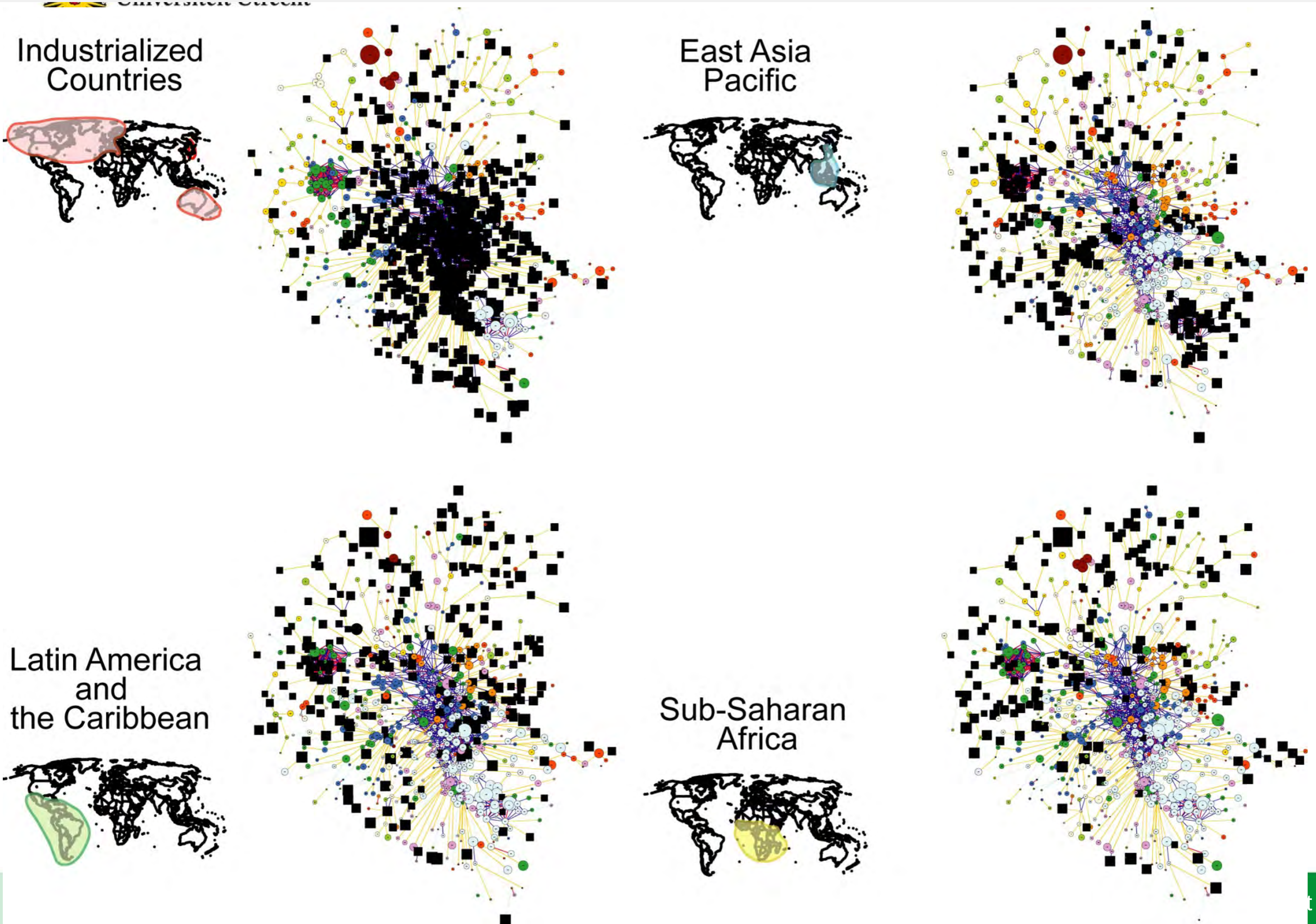
How do countries accumulate capabilities?



CA Hidalgo, B Klinger, A-L Barabasi, R Hausmann.
Science (2007)

Patterns of Comparative Advantage

Hidalgo et a. *Science* (2007)



Data for network analyses

- The story is the same in one field after another, in science, politics, crime prevention, public health, sports and industries as varied as energy and advertising.
- All are being transformed by data-driven discovery and decision-making.
- Anderson claims the end of models and theory in WIRED: The Data Deluge Makes the Scientific Method Obsolete.

Scientometric databases: Web of Science, Scopus and Google Scholar

For years researchers looking for scientometric information had only one resource to consult: the Web of Science from Thomson Scientific.

Using an IP address at Utrecht University, you have access to the *Science Citation Index* via the so-called Web-of-Science. The database can be found in the digital library of the university (http://bibe.library.uu.nl/zoek/biblio/index_e.html).

In 2004 two competitors emerged – Scopus from Elsevier (also available at Utrecht University) and Google Scholar from Google (freely available).

Scientometrics

- Essentially, the science of measuring and analysing science.

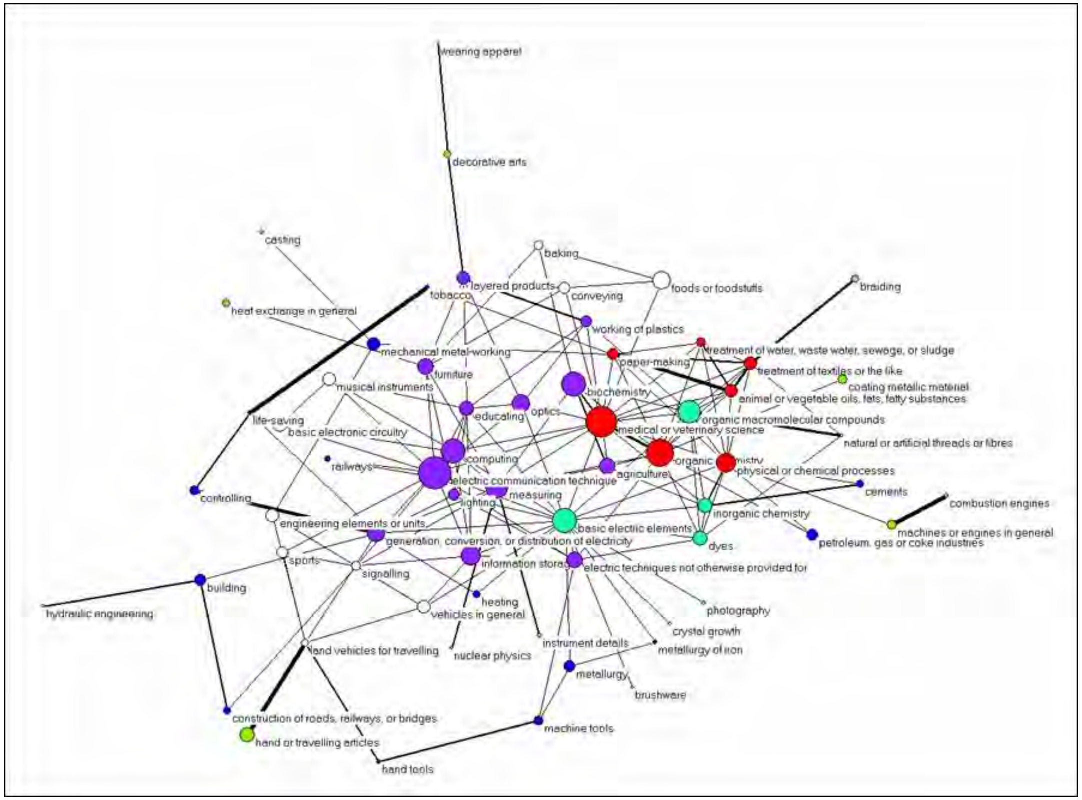
- The unit of analysis
  - can range from an individual to entire countries
  - can be a topic within a scientific field or the whole field
  - can be an object (e.g. a piece of lab equipment) or it can be abstract (e.g. a potential concept to emerge in 10 years from a specific field).

- Core journals
  - Scientometrics
  - JASIST

```
FN ISI Export Format
VR 1.0
PT J
AU Zangirolami, TC
   Johansen, CL
   Nielsen, J
   Jorgensen, SB
TI Simulation of penicillin production in fed-batch cultivations using a
   morphologically structured model
SO BIOTECHNOLOGY AND BIOENGINEERING
LA English
DT Article
DE structured model; fed-batch cultivation; repeated fed-batch
   cultivations; penicillin production; Penicillium chrysogenum
ID MATHEMATICAL-MODEL; SUBMERGED CULTURES; CHRYSOGENUM; FRAGMENTATION;
   FERMENTATION; GROWTH; STRAIN
AB A mathematical model is formulated to describe trends in biomass and
   penicillin formation as well as substrate consumption for fed-batch
   cultivations. The biomass is structured into three morphological
   compartments, and glucose and corn steep liquor are considered as
   substrates for growth. Penicillin formation is assumed to take place
in
   the subapical compartment and in the growing region of the hyphal
   compartment. Furthermore, it is inhibited by glucose. Model parameters
   are estimated using an evolutionary algorithm and fitting the model to
   a standard fed-batch cultivation. The model is validated on
   experimental data from three different fed-batch cultivations,
   including two repeated fed-batch cultivations. The model predictions
   show good agreement with the measurements of biomass and pencillin
   concentrations for all fed-batch cultivations. (C) 1997 John Wiley &
   Sons, Inc.
C1 TECH UNIV DENMARK,DEPT BIOTECHNOL,CTR PROC BIOTECHNOL,DK-2800
LYNGBY,DENMARK.
   UNILEVER RES LABS VLAARDINGEN,NL-3133 AT VLAARDINGEN,NETHERLANDS.
   TECH UNIV DENMARK,DEPT CHEM ENGN,DK-2800 LYNGBY,DENMARK.
CR BAJPAI RK, 1980, J CHEM TECHNOL BIOT, V30, P332
   BAJPAI RK, 1981, BIOTECHNOL BIOENG, V23, P717
   CARLSEN M, 1993, ANAL CHIM ACTA, V279, P51
   CHRISTENSEN LH, 1992, THESIS TU DENMARK LY
   FIDDY C, 1976, J GEN MICROBIOL, V97, P169
   HEIJNEN JJ, 1979, BIOTECHNOL BIOENG, V21, P2175
   HOLMBERG A, 1982, MATH BIOSCI, V62, P23
   JOHANSEN CL, 1993, THESIS TU DENMARK LY
   JORGENSEN H, 1995, APPL MICROBIOL BIOT, V43, P123
   JORGENSEN H, 1995, BIOTECHNOL BIOENG, V46, P117
   MEGEE RD, 1970, BIOTECHNOL BIOENG, V12, P771
   MENEZES JC, 1994, J CHEM TECHNOL BIOT, V61, P123
   NESTAAS E, 1983, BIOTECHNOL BIOENG, V25, P781
   NICOLAI BM, 1991, BIOTECHNOL LETT, V13, P489
   NIELSEN J, 1992, ADV BIOCHEM ENG BIOT, V46, P187
   NIELSEN J, 1993, BIOTECHNOL BIOENG, V41, P715
   NIELSEN J, 1995, BIOTECHNOL BIOENG, V46, P588
   NIELSEN J, 1995, BIOTECHNOL PROGR, V11, P93
   NIELSEN J, 1995, THESIS TU DENMARK LY
   PAUL GC, 1994, BIOTECHNOL BIOENG, V44, P655
   PAUL GC, 1996, IN PRESS BIOTECHNOL
   PISSARRA PN, 1995, IN PRESS BIOTECHNOL
   SCHMIDT K, 1995, P NORD MATL C STOCKH
```

Pajek

Pajek (Slovene word for Spider) is a program, for Windows, for analysis and visualization of large networks. It is freely available, for noncommercial use, at its webpage.


wearing apparel
decorative arts
casting
baking
layered products
tobacco
conveying
foods or foodstuffs
braiding
heat exchange in general
working of plastics
mechanical metal-working
treatment of water, waste water, sewage, or sludge
paper-making
furniture
treatment of textiles or the like
musical instruments
biochemistry
coating metallic material
animal or vegetable oils, fats, fatty substances
educating
optics
organic macromolecular compounds
life-saving
medical or veterinary science
basic electronic circuitry
natural or artificial threads or fibres
computing
organic
mistry
railways
physical or chemical processes
agriculture
electric communication technique
cements
controlling
lighting
measuring
combustion engines
inorganic chemistry
engineering elements or units
basic electric elements
machines or engines in general
generation, conversion, or distribution of electricity
dyes
petroleum, gas or coke industries
sports
information storag
electric techniques not otherwise provided for
signalling
building
heating
photography
hydraulic engineering
vehicles in general
crystal growth
instrument details
metallurgy of iron
land vehicles for travelling
nuclear physics
metallurgy
brushware
construction of roads, railways, or bridges
machine tools
hand or travelling articles
hand tools

Gephi

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

Runs on Windows, Linux and Mac OS X. Gephi is open-source and free

SAINT

- The Rathenau Institute has developed SAINT, which stands for Science Assessment Integrated Network Toolkit. This is a set of tools for bibliometric and patentometric research. This toolkit can be downloaded from this website. In return, we would like your help.

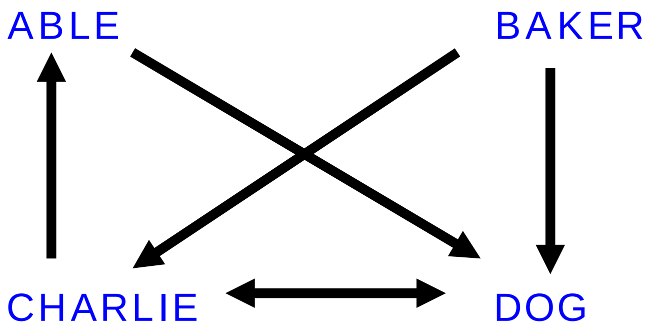
- Please note that MS Access is required for this tool.

There are three main reasons for using "formal" methods in representing social network data:

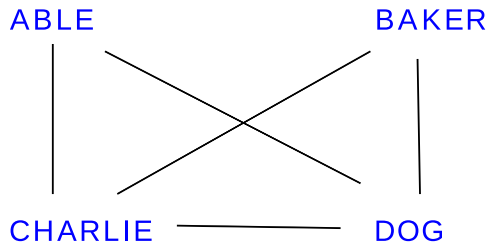
- Matrices and graphs are compact and systematic: They summarize and present a lot of information quickly and easily
- Matrices and graphs allow us to apply computers to analyzing data
- Matrices and graphs have rules and conventions: Sometimes they lead us to see things in our data that might not have occurred to us to look for if we had described our data only with words.

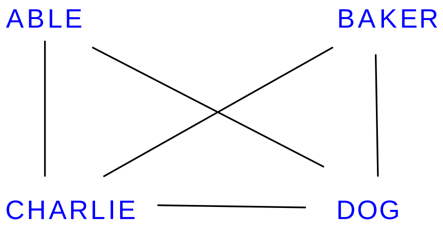
Data

- Network can be presented on input file in different ways:
  - using arcs/edges (e.g. 1 2 – line from 1 to 2)
  - using arcslists/edgeslists (e.g. 1 2 3 – line from 1 to 2 and from 1 to 3)
  - matrix format



Undirected network





ABLE

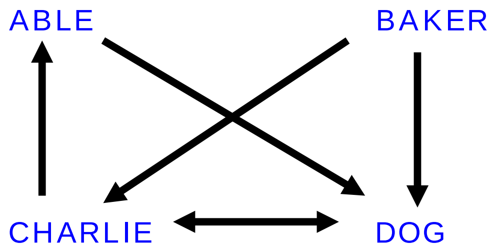
BAKER

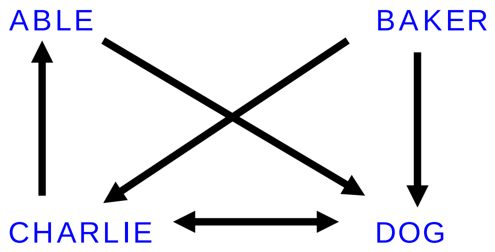
CHARLIE

DOG

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0 & 1 & 1 \\ B & 0 & 0 & 1 & 1 \\ C & 1 & 1 & 0 & 1 \\ D & 1 & 1 & 1 & 0 \end{array}$$

Directed network





$$
\begin{array}{ccccc}
 & A & B & C & D \\
A & 0 & 0 & 0 & 1 \\
B & 0 & 0 & 1 & 1 \\
C & 1 & 0 & 0 & 1 \\
D & 0 & 0 & 1 & 0
\end{array}
$$

- **fullmatrix** input (one-mode), sending actors appear in rows and receiving actors in columns (in the identical order).
- The “dl” line states the number of actors and the input format.

```
dl n=4 format=fullmatrix
labels:
able baker charlie dog
data:
0 0 0 1
0 0 1 1
1 0 0 1
0 0 1 0
```

In **nodelist1** input (one-mode), the first number in each data line is the row of the sending actor, followed by the numbers of the actors to which it sends a relation.

```
dl n=4 format=nodelist1
labels:
able,baker,charlie,dog
data:
1 4
2 3 4
3 1 4
4 3
```

2 mode networks

- 2-Mode Matrix. A (2-dimensional) matrix is said to be 2-mode if the rows and columns index different sets of entities (e.g., the rows might correspond to persons while the columns correspond to organizations).
- In contrast, a matrix is 1-mode if the rows and columns refer to the same set of entities, such as a city-by-city matrix if distances.

- To input **two-mode** matrices, the dl and labels commands specify distinct row and column units. For example:

```
dl nr=45 nc=12 format=fullmatrix
row labels:
able baker charlie dog easy fox . . . . .
column labels:
january,february,march,april, . . . .
```

Relations: Content based

- The value for each term in a document vector can be chosen based on the application.
- Co-word analysis leaves the level of individual documents and creates pairs of keywords or terms.

2-mode

	Word1	Word2	Word3	….	….	Word n
Text 1						
Text 2						
Text 3						
…						
…						
…						
Text n						

1-mode

	Word1	Word2	Word3	….	….	Word n
Word1						
Word2						
Word3						
…						
…						
…						
Word n						

- A lot of the work that we do with social networks is primarily descriptive and/or exploratory, rather than confirmatory hypothesis testing.
- Using network tools can be particularly helpful because they may let you see patterns that you might not otherwise have seen.
- The tools can be used to explore tentative empirical generalizations and provide crude first examinations of hypotheses about patterns that may be present in the data.