IMT SCHOOL FOR ADVANCED STUDIES LUCCA

Data Science & Statistical Learning | II Level Master

Network and Media Analysis

UNIVERSITÀ DEGLI STUDI FIRENZE

How the Taxonomy of Products Drives the Economic Development of Countries

Prof. *Tiziano Squartini*

Gianmarco Santoro

22/03/2024

- **Introduced** an **algorithm**: **reconstruct network** structure of **time-evolution country-product bipartite networks**
  - **Links** (edges) are obtained by **selecting largest values of projected matrix**
- **Exports time series give** a fundamental **insight to understand countries' development**:
  - **Supposing that products are defined by capabilities needed to produce it**, presence of a product in a country's export basket represent a hint on capability basket of country itself
  - **One can build a network of products in which two products are connected if they share some capabilities**, so it's possible to avoid studying capabilities structure of countries, which is very hard to quantify
- **This network** of products, **introduces for a different approach**:
  - **Two nodes are connected by a directed link which represents causality relationship between them**:
    - ***a*** and ***b*** are connected if one of them, say ***a***, **makes more probable that *b* will be produced in future**:
      - In this case, a directed link will go from ***a*** to ***b***
- Proposed **algorithm**:
  - **Filters info contained in empirical export data**
  - **Builds hierarchical network**: **nodes are products** and **directed links are necessary relationship** between products
  - **Reduces number of edges from almost fully connected projection of bipartite country-product network**:
    - **~ $N^2 \rightarrow N$** selecting **most informative ones in economic progress**

Products Taxonomy | Data

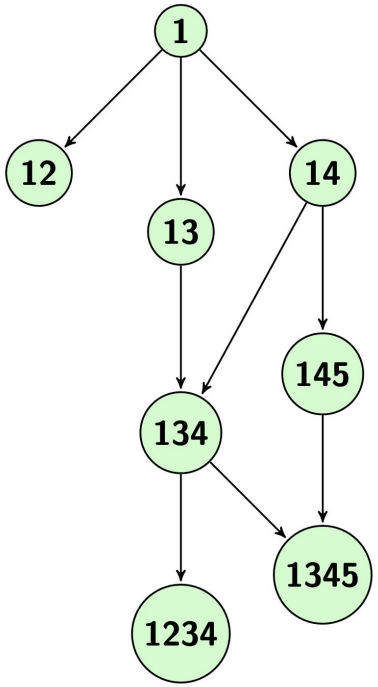
- **2 databases**, collected by United Nations, reporting **import-export flows between countries**:
  - Number of **countries** ranges **from 134 to 151**
- Cleaning process to **remove clear errors and inconsistencies**
- Build a matrix $\mathbf{M}_{cp}$: elements are equal to **1 if country *c* exports product *p*** and **0** otherwise
- Values are assigned using a threshold on ***Revealed Comparative Advantage*** defined by ***Balassa***

- Each year has a defined import-export structure, so resulting matrices are different
- Databases have a different number of products, categorized with different classifications
- So, during cleaning process number of products in each databases is kept constant through years

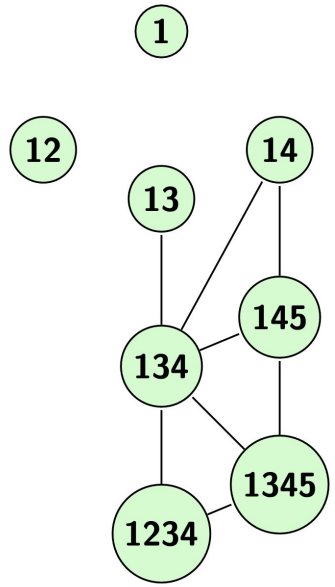
- Build and analyze **two networks**:
  - **1st** referring to years **1995-2010**, contains **1131 products** classified in HS2007
  - **2nd** spanning from **1963 to 2000**, has a lower number of **products (538)**, classified in SITC rev.4
    - Latter permits an analysis of development of countries on a longer time horizon, on several economic cycles

Products Taxonomy | Taxonomy and Proximity

- Build a hierarchically ordered network: **structure is inferred from M_{cp}** matrix
- **Products defined in terms of capabilities** needed **to produce them**:
  - E.g. capability **1** corresponds to a basic product. A country equipped with a second capability **2**, can export "**12**" product. Capabilities **1, 2** and **3** could simply not lead to a product, while "**134**" can be a product, and so on
  - A **hierarchy naturally arises**: **some products are mandatory intermediate steps to be able to produce more complex technologies** and **sons are connected to father by a directed edge**
  - Figure **(a)** shows an example of this structure, so called ***taxonomy network***
  - Figure **(b)**: a ***proximity network***, same **products are connected if they share a fraction of their composing capabilities**. It has an **undirected network**, because **products are connected if they are similar, so at same level**
- A **country** will likely **move from basic products to more complex** ones **when it develops new capabilities**: time evolution of **technological progress should be closer to a *taxonomy* than to a *proximity* network**



(a) Taxonomy Network



(b) Proximity Network

Connect same products, characterized by capabilities needed to produce them

On left, a hierarchical relationship

On right, joined similar products

- d_c define ***diversification*** as **number of products exported by country c**, as measured by *Revealed Comparative Advantage*:

$$d_c = \sum_p M_{cp}$$

- ***Ubiquity***, u_p is **number of countries which export product *p***:

$$u_p = \sum_c M_{cp}$$

- M_{cp} **projected** to obtain a **product-product matrix of normalized probability**:

$$B_{pp'} = \frac{1}{\max(u_p, u_{p'})} \sum_c \frac{M_{cp} M_{cp'}}{\sqrt{d_c}}$$

- A way to **normalize projection**: $\sqrt{d_c}$ **factor takes into account different contribution given by countries of different diversifications**, by dividing corresponding terms by expected values in a random binomial case
- **To obtain adjacency matrix** of a network with **number of edges of same order of magnitude of number of products selected only maximum entry of each row, excluding diagonal elements**:
  - For each product p look for product $p' \neq p$ **which maximizes normalized probability** $B_{pp'}$ **to be exported in a pair**

- **Selected product contributes most with respect to its column**: pick product whose column has smallest elements. **This filtering procedure discards** redundant and **noisy info and define** a set of preferred **patterns for development policies**
- 38 different matrices 1st dataset and 16 in 1nd: **aggregate years in a single matrix with same columns** of exported products and, as rows all countries, including repetitions due to different years: this way most of fluctuations are averaged out
- ***Product Space*** is based on **proximity** $\varphi_{pp'}$ between products p and p', defined as:

$$\phi_{pp'} = \min\left(\frac{\sum_c M_{cp}M_{cp'}}{u_p}, \frac{\sum_c M_{cp}M_{cp'}}{u_{p'}}\right)$$

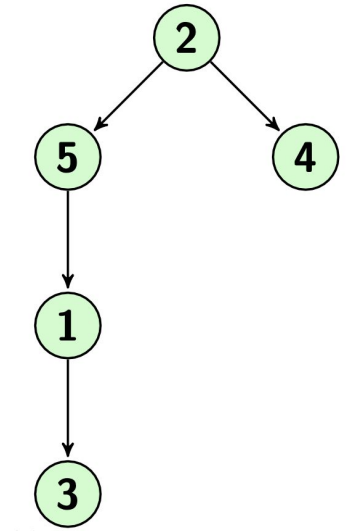
- When used without any further filtering process, **leads to an almost complete weighted network**. Purpose of maximum picking procedure is to enhance signal to noise ratio in such a way to build a conceptually different network, whose links are directed and related to necessity instead of proximity
- **Differences between Taxonomy Network** and **Product Space**:
  I. Presence of **directed links**, with a **causality meaning**
  II. **Link number reduction** of order **~ $N^2 \rightarrow N$**
  III. **Different normalization** considering different **diversifications or countries**

Products Taxonomy | Tests

- Algorithm output example: **starting from a simple M_{cp}**, figure shows **matrix and resulting taxonomy network**
- Countries in rows and products in columns; e.g. 2nd country produces 2nd and 5th product
- Focusing on relationship between structure of matrix and of network:
  - **Product 2**, 2nd column, is **made by all countries**: this means that, probably, **capabilities needed are few or simple**
  - **Products 3 and 4 are exported by only one country**, so **very specific features are required** by these products
  - Products **5 and 1 lay** somehow **in middle**
- **Ubiquitous product 2 results to be root**, and **it is needed to make all other products**:
  - **Country 4 exports** only **products 2 and 4** suggests that **capabilities to produce 2 are mandatory to produce 4**
  - Left branch is a chain of products built following same reasoning

$$\begin{array}{c} \\ \text{countries} \end{array} \begin{array}{c} \text{products} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

(a) M_{cp} matrix



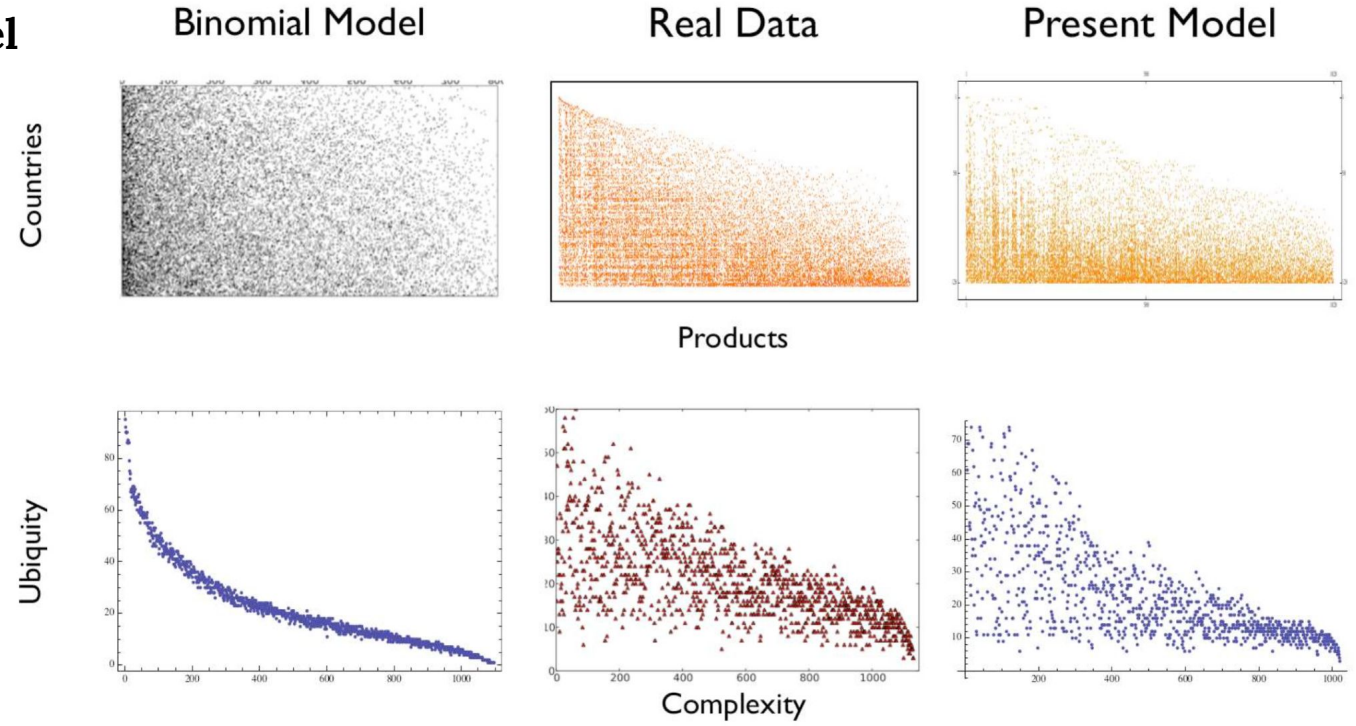
(b) Taxonomy Product Network

Products Taxonomy | Tests

- Built a **model to test**: **known relation between products** to obtain taxonomy, $\mathbf{M_{cp}}$ and **compare reconstruction** methods
- Model: construction of product taxonomy starts with ***R* root products**:
  - These **products needs only one capability** in order **to be produced**
  - A **capability** is **minimal and non-trivial endowments** needed in order to produce a product
    - **Non-trivial**: **not owned by all countries** by default (a trivial capability could be water or sunlight)
    - **Minimal**: a capability **is smallest set of endowments to produce a new product** in at least one case
- Product **taxonomy building**:
  1. **At each time step a new capability is introduced**
  2. **New capability defines new product p'** being **added to one of existing products p** with **uniform probability**
  3. **A directed link is inserted from p** to **p'**
- Then $\mathbf{M_{cp}}$ **matrix is built** as follows:
  1. A **diversification d_c** is **assigned to each country *c***; specific value is extracted from a real-world distribution
  2. Country chooses randomly d_c **products from taxonomy**, **probability of choosing a particular product is inversely proportional to number of capabilities** (**distance from root**) associated with that product
  3. **Products on shortest path from root** of corresponding tree of any chosen product **are assigned to country *c***

Products Taxonomy | Tests

- Values of d_c are chosen such that distribution of diversification in model is similar to the one coming from real data
- Model is able to reproduce some non-trivial stylized facts present in real $\mathbf{M}_{cp}$ matrix
- **Figure shows a simple binomial model**, **no product taxonomy** is present, **real $\mathbf{M}_{cp}$** matrix **and a realization of this model**
- First row shows a representation of matrices: **binomial model misses some aspects**, while **taxonomy based model produces results closer to real case**; confirmed looking at scatter plot of ubiquity vs complexity ranking of products
- **Complexity is a measure of number of needed capabilities**: **triangular shape** (**existence of products that are both ubiquitous and complex**) present in real-world data is very difficult to reproduce with even more complicated binomial models but **emerges naturally from model**



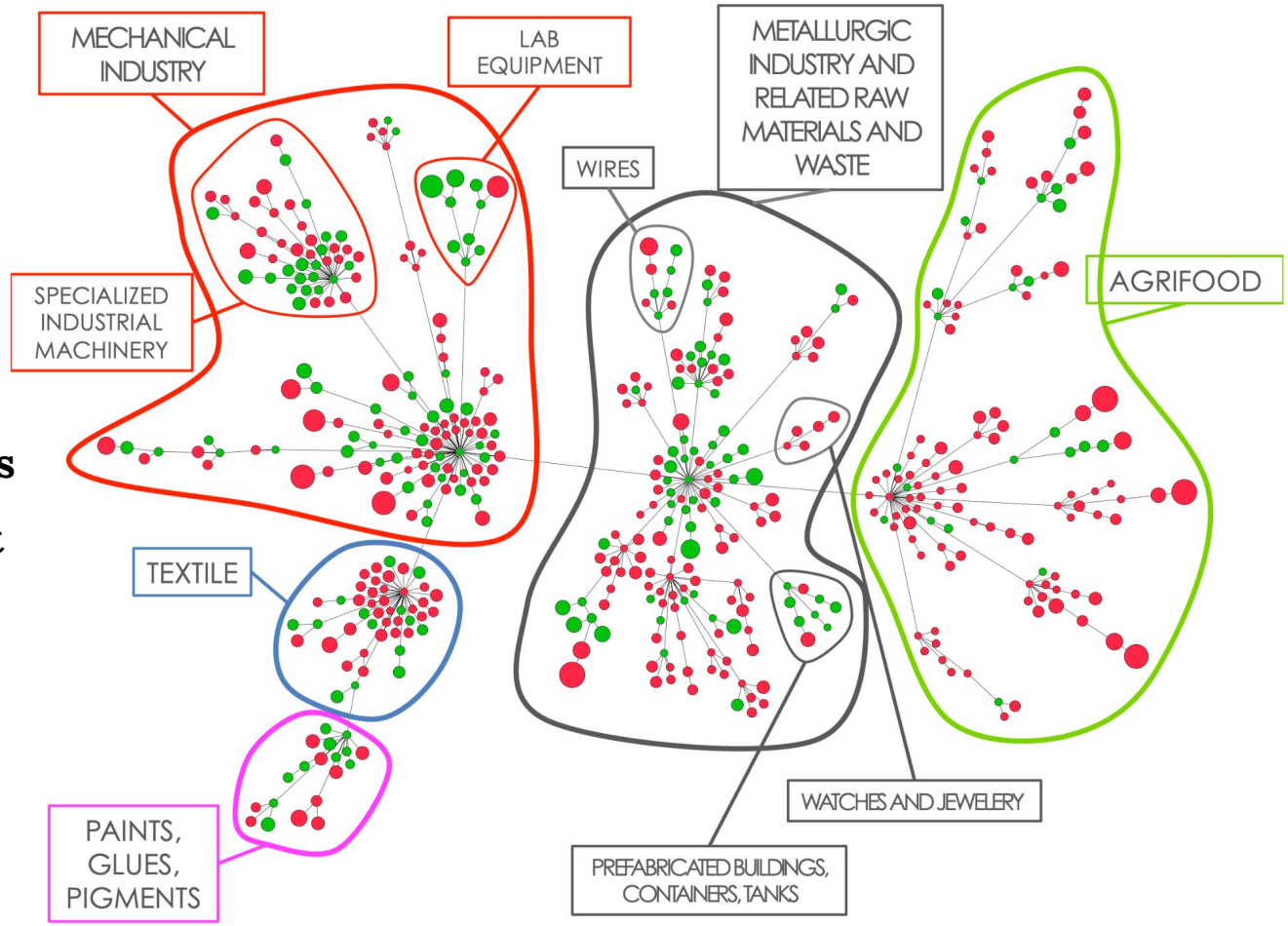
- This model is used to benchmark performance of various algorithms in reconstructing taxonomy by starting from $\mathbf{M}_{cp}$ matrix: **with 120 countries and 1120 products, algorithm detects more of 80% of correct links**
- Comparison among three different ways to reconstruct a taxonomy network:
  - Present algorithm outperforms random assignment of link and maximum spanning tree obtained from same matrix

% of correctly reconstructed links		
	25 countries 248 products	120 countries 1120 products
Present Criteria	42.4%	81.1%
Max. Spanning Tree	25.8%	33.4%
Random links	10.3%	11.3%

- **Results**
  - **2 taxonomy networks** built starting **from empirical data**:
    - Network obtained from 1995-2000 data has 1131 vertices (equal to number of products) and 985 edges
    - 1963-2000 network has 538 vertices and 456 edges
  - **Both quite sparse and not fully connected**, due intrinsic products heterogeneity and to filtering, **selecting at most one link per row permits to identify most relevant links from point of view of observed time evolution**
  - Both with about 100 components with heterogeneous sizes, most of these components have a well defined economical and technological meaning

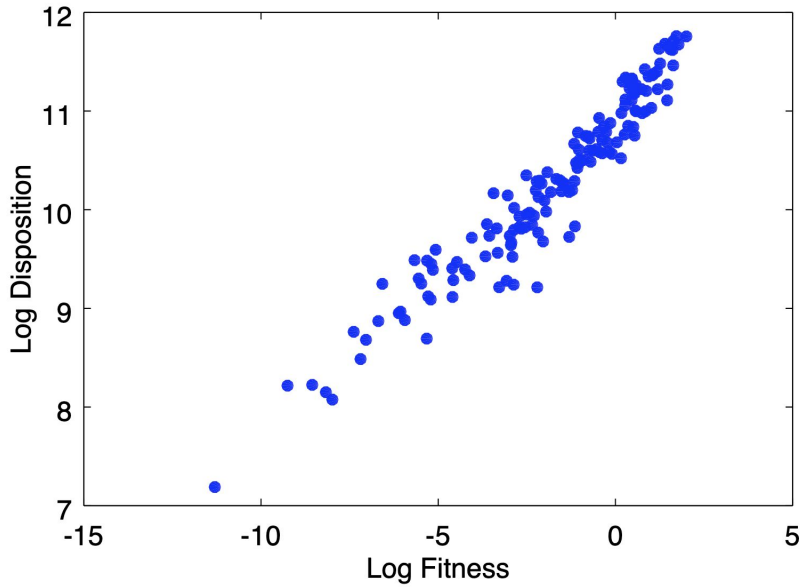
Products Taxonomy | Results

- **Figure shows largest component of taxonomy network** built **from 1995-2010 export matrices**
- **Green filled nodes** (= 1) represent **products that are exported by Sweden in 2010**, while **red ones have $M_{cp} = 0$**
- **Diameter** of vertices is **proportional to logarithm of product complexity**
- From a visual inspection, a good strategy for Sweden could be to produce red, high complexity product in Lab Equipment community
- There's **tendency for products of large complexity to be on border of network**, while **more basic products lay in center and have a higher degree**:
  - **Centrality tends to be anticorrelated with complexity**
- This behavior **confirms hypothesis** that **few capabilities needed to produce low complexity products represent a necessary condition to be able to produce high complexity products**, according to Taxonomy concept
- A validation: **about 70% of edges point from a low to a high complexity product**



- **Developed countries tend to occupy outlying vertices**, to measure it, **used centrality of a given vertex** which takes into account not only **its degree but also direction of links**, **to pass received authority following links**
- **One possible measure is *PageRank***: evaluate **degree of development** of a country, **counting its products weighting more if lie away from center, which are vertices with a low PageRank**
- **Sum of inverse PageRank of exported products** of a given country ***c***:

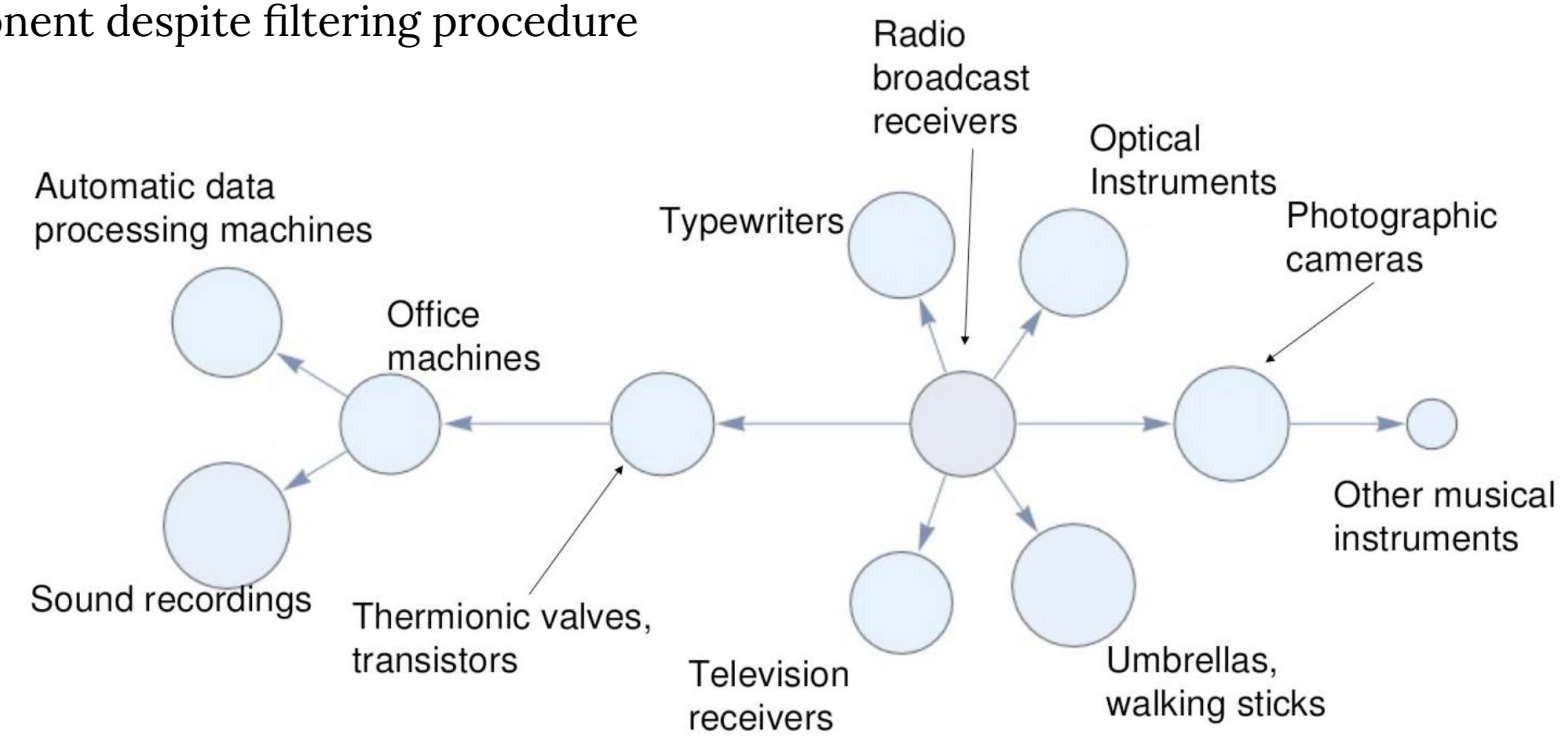
$$D_c = \sum_p M_{cp} PR_p^{-1}$$



- **Called *disposition* of country**: **plotted versus fitness**, both log, **that is a measure of growth potential** of a country, referring to year 2000, **finding high correlation** with $\mathbf{R^2 = 0.92}$
- This connection between a network based quantity and fitness is result of algorithmic interplay between countries and complexity of products they export, so **indicating a connection between growth potential of a country and its disposition on taxonomy network**

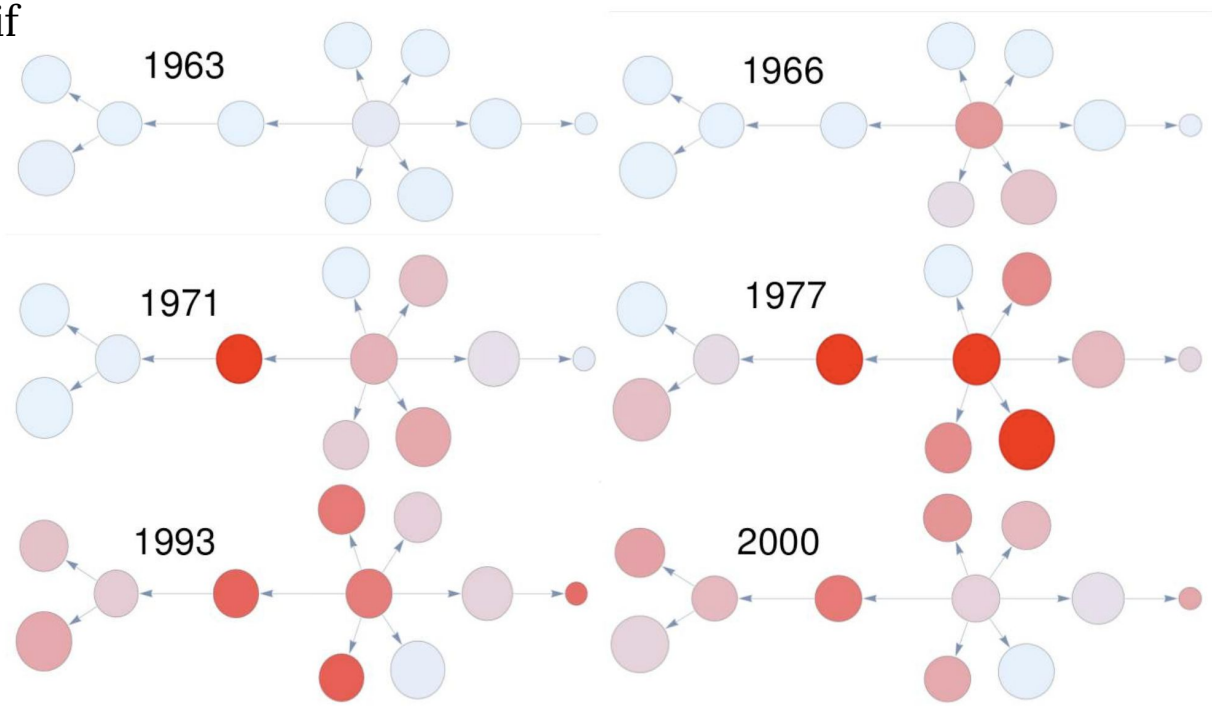
Products Taxonomy | Results

- One of most **important features of this approach is visualization** of countries' economic development
- **Patterns emerge when studying specific countries through time**: example of **development of South Korea**, often reported as a case study for a successful industrialization process
- **Figure shows tech-component of taxonomy network**:
  - **Root product is radio broadcast receivers**, while **on border**: automatic data processing machines → **computers**
  - An **evident exception is umbrellas**, **due to noise**, a product with **nothing in common with others**, still connected to this component despite filtering procedure



Products Taxonomy | Results

- **Figure shows time-evolution-South-Korean-export**, colors are proportional to *Revealed Comparative Advantage* (RCA):
  - **Light blue means product is not exported**, while **different shades of red are proportional to an RCA increase**
- **In 1963** this country **did not export any product of this component** in a significant way. **After 3 years**, **root starts to be produced together two close products**. **In following years** South Korea **explores network**, reaching **in 1993 an impressive level of diversification**. **In 2000** South Korea **focused its exports on borderline products**, **as expected from** an already **developed country** from disposition analysis presented above
- **Presence of** a meddlesome product (in this case, **umbrellas**) is **due to noise**, but it **can be spotted thanks to its RCA** behavior, **which is uncorrelated with other nodes**. So, even if probabilistic approach used to define network can lead to spurious results, like presence of unexpected products in otherwise well defined clusters, one can see that a careful analysis of dynamics clearly points to fact that this site is anomalous with respect to cluster considered
- One can **notice diffusion from center (root product) towards borders of component**



- **Use this network structure** to quantify on a larger scale what observed in South Korean case, via (1963-2000) database and **trying to extract**, from export matrices, empirical info about **correlation between presence of a product in a country's export basket and appearance of a new one in a future year**
- Quantified how much presence of a product ***p*** influences possible turning on of another product ***p′***. **One possible measure** of this helpfulness **is frequency of activations given presence of an already activated product**. In practice, calculate three dimensional ***Activation Matrix***:

$$Z_{cpy} = M_{cpy} - M_{cp(y-1)}$$

- Where **y ∈ [1964, 2000]**. **Focusing only on activations of products (*Zcpy* = 1)**, ignoring ***Zcpy* = −1**, which corresponds to dismission of a certain production. To evaluate **frequency of activation** of ***p′*** given presence of another product ***p***, calculate ***Enabling Matrix***:

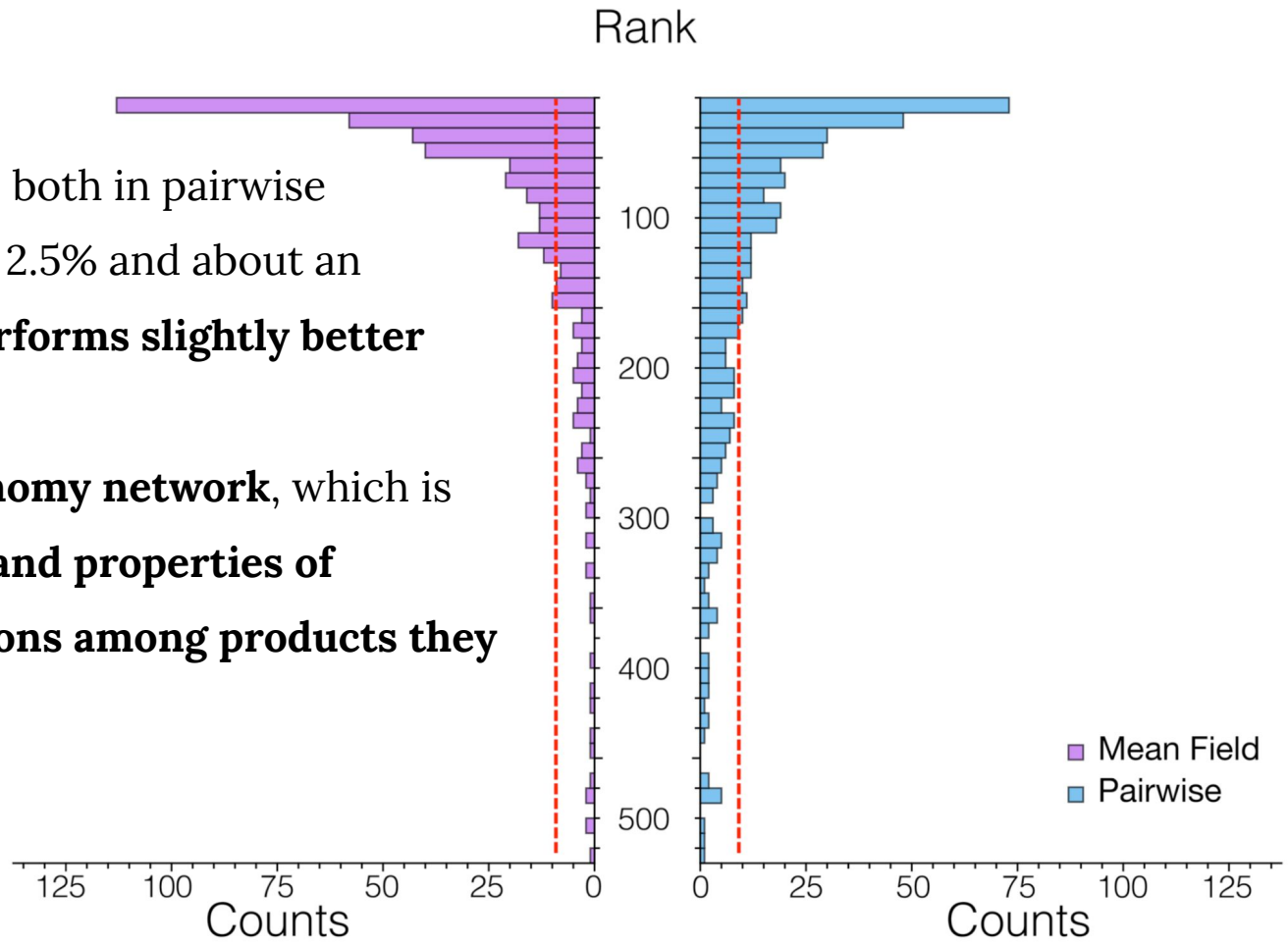
$$C_{pp'} = \frac{\sum_{c,y} Z_{cp'y} M_{cp(y-1)}}{\sum_{c,y} Z_{cp'y}}$$

- Where matrix operations are intended as element by element operations
- **Elements of this matrix represent an empirical estimation of strength of directed link from *p* to *p′***

- It's an approximation: in principle more a product is present, more it will appear to be necessary even if it could be not
  - Checked weight of products' ubiquity, finding that, even if ubiquitous products tend to be more necessary, once that this effect is removed results are substantially left unchanged
- **Another possibility** is that **the presence of a set of products changes probability that a country has to produce a new product, and not only one as supposed above**
  - It's hard to calculate **relative usefulness of products** so, as an approximation, **give for every activation a score $1/n$ to each product which was already exported during previous year**, where n **is number of products exported by country**, so a function of p, c and y
  - With this approach **empirical strength of link from p** to p' **is given by sum of scores collected by different countries through years**
- **In this way, Enabling Matrix is calculated supposing a *mean field interaction*:**
  - **In contrast with previous approach, in which interaction was assumed to be *pairwise***

Products Taxonomy | Results

- **Once defined an empirical benchmark regarding time evolution of countries' export**, assessed its connection with taxonomy network: sorted rows of two enabling matrices from largest to smallest element and checked position of matrix element that would have picked following taxonomy network
  - Practically, **checking how strong is empirical realization of link present in Taxonomy Network with respect to other possible links**
  - Taxonomy network correctly identifies most of top empirical temporal connections between products, both in pairwise and in mean field approach. ~100 of links are in top 2.5% and about an half of them are in top 10%. **Taxonomy network performs slightly better in mean field case**
- This result points out a clear **connection between taxonomy network**, which is built up without considering time evolution of exports, **and properties of countries' development in terms of temporal connections among products they are exporting and ones they will export**



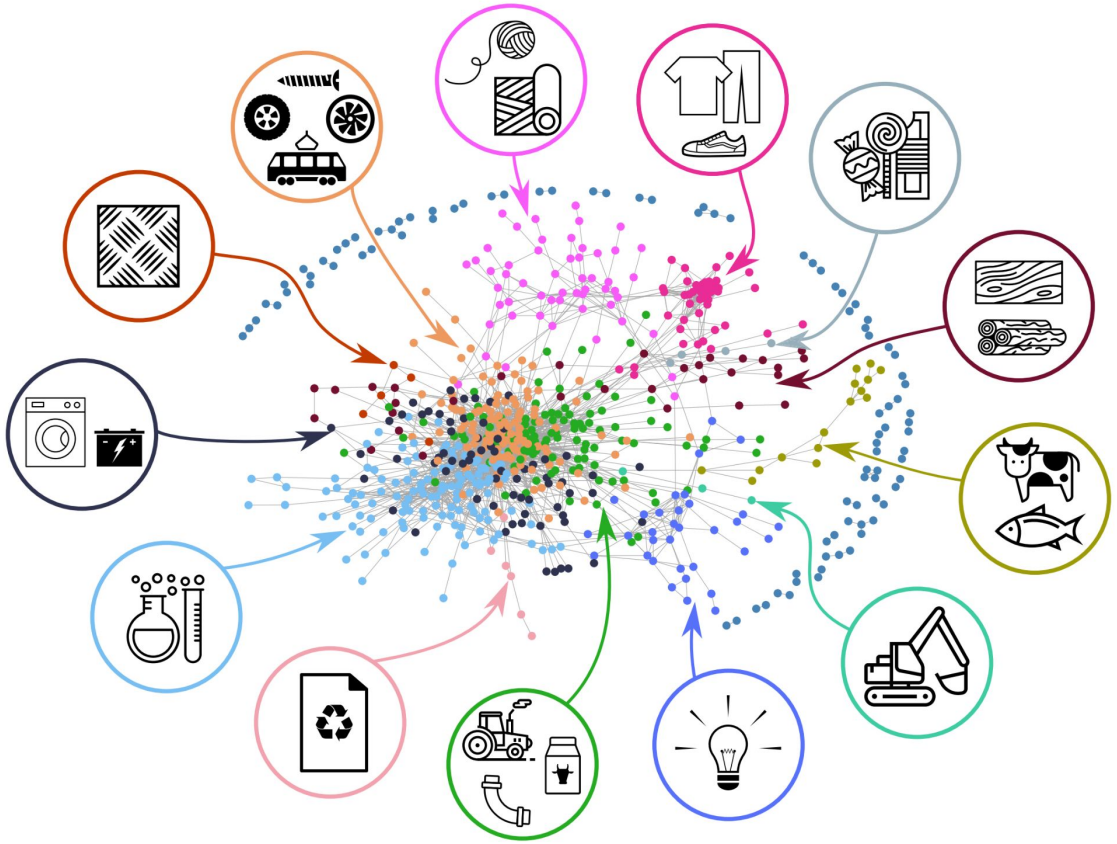
Products Taxonomy | Discussion

- This work introduces an **algorithm to extract relevant info from time evolution of a bipartite network**
  - **Build a directed network** whose nodes are constituted by only **one typology of nodes of starting bipartite network** and whose **edges point from a required node to a supported node**:
    - In sense that **activation of first node increases probability that second node will be activated in future**
  - **Given this causality relationship, named this a *taxonomy network***
- **Algorithm, based on picking maximums of projection of bipartite network**, is tested on simple matrices and with a toy model which is able to reproduce main stylized facts of export matrices
- **Used this framework to analyze taxonomy network** resulting **from export data** of countries:
  - **Network properties are connected to countries' potential growth and development**:
    - Last aspect is investigated by introducing **enabling matrix**, whose **elements are a measure of how necessary is an activated product to be able to produce another product in future**
    - Found that largest temporal connections **from an empirical analysis of export matrices are the ones would have picked by looking at structure of taxonomy network**
    - This fact links static properties of taxonomy network with time evolution of bipartite one

- This work opens up possibility to a number of possible applications:
  - **In general, this algorithm could be used in any bipartite system, especially in cases in which one topology of nodes play an active role in choosing which node pick from other topology**, **for example** in country-product networks, user-item, consumer-purchased product, and **in all other recommendation systems**
- About country-product network, **next step would be link prediction**: **build a framework able to predict which product will be exported by a given country in next years**:
  - For example, **one could look for products linked to ones which are already exported by country** in analysis
    - With this framework: fewer capabilities are needed to make that step. Then **taxonomy network can be used to give policy suggestions, because a product which is close to many already produced is easier to produce**
    - Correspondence between this network structure and empirical time connections measured by Enabling Matrix suggests that **there is a well defined path to follow in industrialization process**:
      - In product space possible trajectories are many, but a number of them are preferred ones to achieve countries' growth. Simply copy other countries without learning their capabilities cannot give long lasting results in terms of enduring economic stability
      - **A less developed country has to learn simple capabilities and to be consequently able to export so called root products in order to start a stable industrialization and development process**

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

UNIVERSITÀ DEGLI STUDI FIRENZE



https://lims.ac.uk/documents/paper-grand-canonical-validation-of-the-bipartite-international-trade-network.pdf

Thanks for your Attention

Images and text have been gathered from the paper: "*How the Taxonomy of Products Drives the Economic Development of Countries*" | *Andrea Zaccaria, Matthieu Cristelli, Andrea Tacchella, Luciano Pietronero | ISC-CNR, Italy; Sapienza Università di Roma, Italy; LIMS, London Institute for Mathematical Sciences, UK | 9 August 2014*"