Consiglio Nazionale delle Ricerche

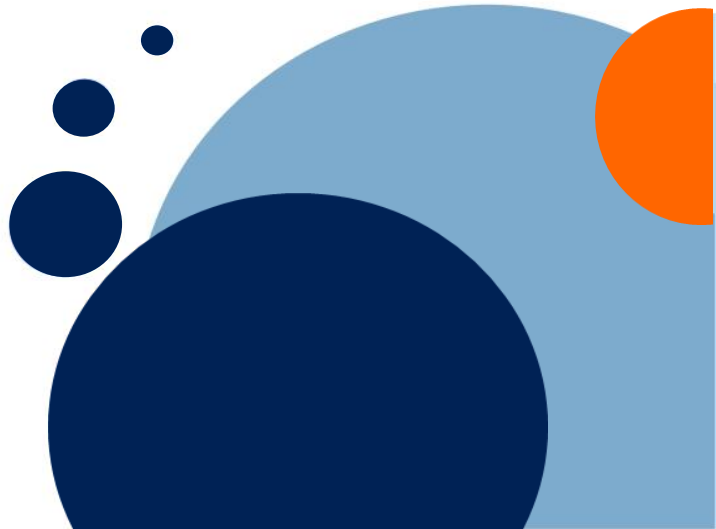
IRCrES

ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
RESEARCH INSTITUTE on SUSTAINABLE ECONOMIC GROWTH

Link prediction in knowledge networks using exogenous and endogenous attributes: a machine learning approach

Antonio Zinilli and Giovanni Cerulli
CNR-IRCrES

Contacts:
antonio.zinilli@ircres.cnr.it
giovanni.cerulli@ircres.cnr.it



Aim of the paper

The aim is to study the differential role played by **nodes' network and non-network attributes** for predicting the collaboration in joint projects of European universities over the time span 2014-2016, in three European Research Council (ERC) domains:

- **Social Sciences and Humanities (SSH)**
- **Physical and Engineering Sciences (PE)**
- **Life Sciences (LS)**

Motivation

The formation (prediction) of prospective links among nodes, is currently one of the most promising research areas of the science of networks (Liben-Nowell and Kleinberg, 2007; Lu and Zhou, 2011; Cho and Yu, 2018; Lande et al., 2020)

By means of recent developments in **machine learning** predictive algorithms, we attempt to estimate the **probability that a university A collaborates with a university B** by considering their idiosyncratic attributes, as well as the past centrality and the sharing of common neighbors

Research questions

The paper addresses three research questions:

1. can collaborations in joint projects be accurately **predicted by machine learning**?

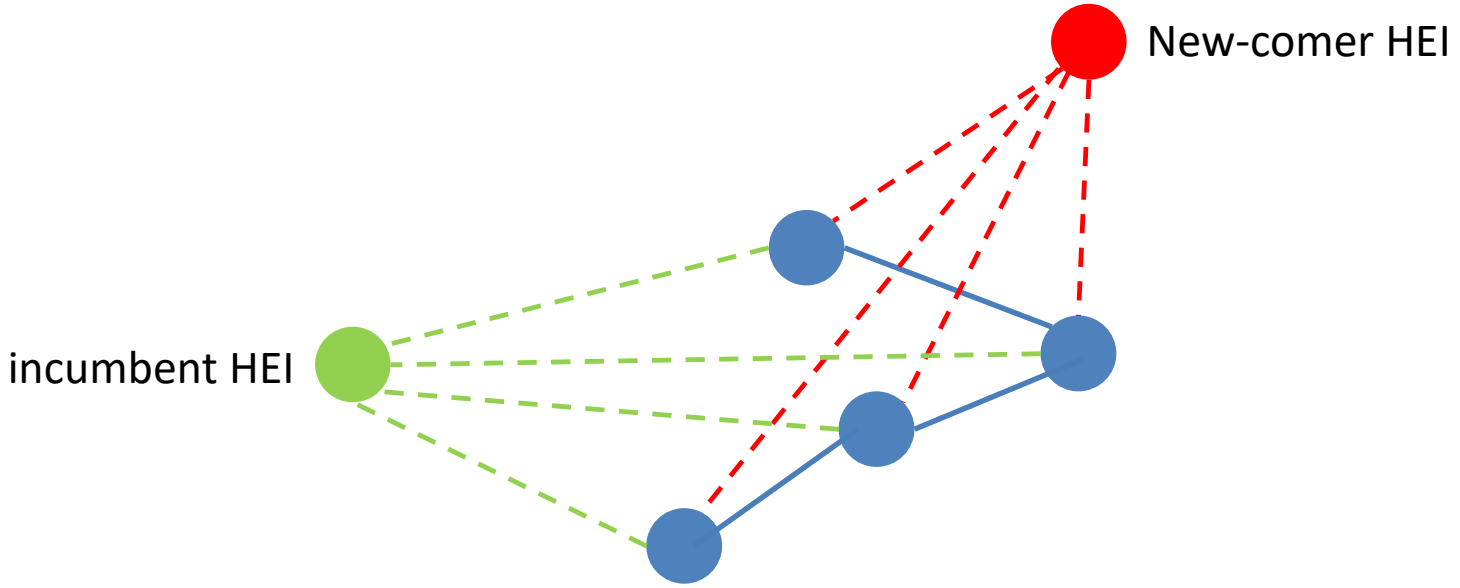
2. which is the predicting power of **endogenous** (*network*) and **exogenous** (*non-network*) characteristics of a node?

3. what features have **larger impact** in predicting links, and in what direction do they act?

Importance of exogenous and endogenous attributes of the node

Incumbent: HEI already part of the network

New-comer: HEI not part of the network yet



For an incumbent HEI, we have both **endogenous** (*network*) and **exogenous** (*non-network*) information

For a new-comer HEI, we only know **exogenous** (*non-network*) information

What is the impact of this on link predictability ?

→ Central to predict accurately **network evolution**

Literature

We look at the **link prediction problem** with the double lens of the network theory (Ahmad et al., 2020; Sun et al., 2020) as our theoretical background, and machine learning (Wang et al., 2015; Nickel et al., 2016) as our data driven approach.

Link prediction is in fact of the utmost relevance in several knowledge network subfields, including co-authorship networks (Shi et al., 2015; Cho et al., 2018; Lande et al., 2020), and future scientific impact of scholars (Hirsch 2007; Mistele et al., 2019).

Analyzing citation and co-publication networks by machine learning techniques, Shibata et al. (2012) concluded that the Jaccard coefficient, the betweenness centrality, and the cosine similarity are powerful factors affecting link prediction.

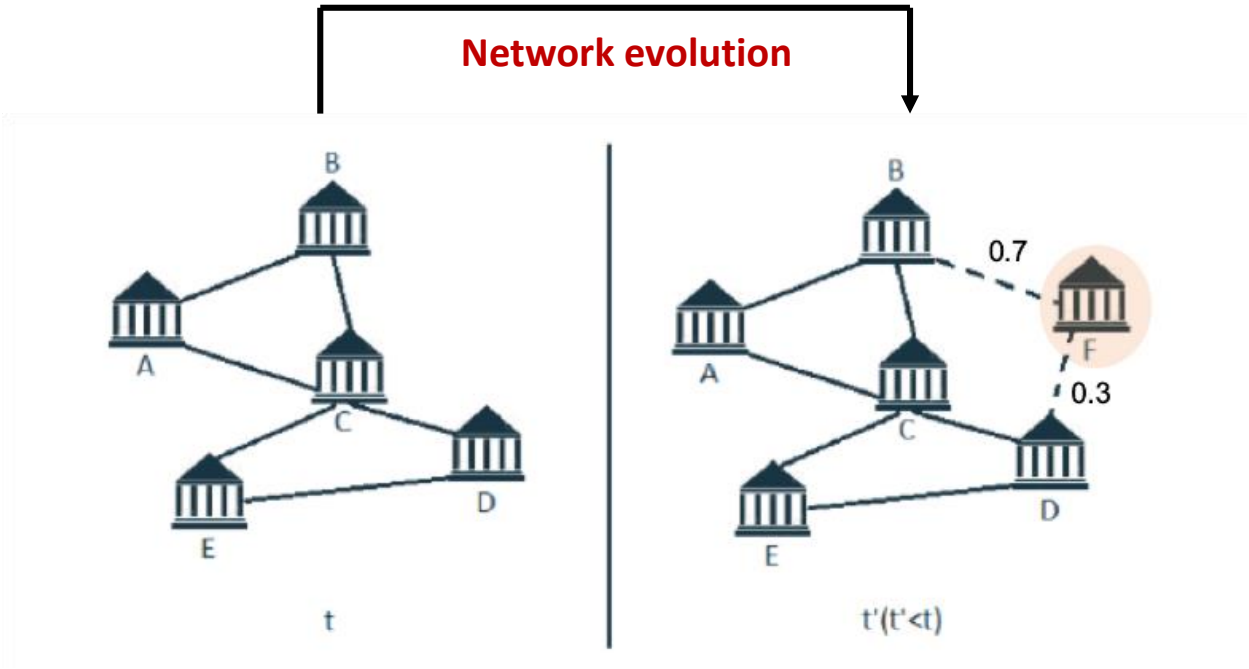
Literature

Many university features play an important role in the collaboration behavior. The **size of a university** is certainly relevant, as larger universities tend to attract more requests for collaborations (Lepori et al. 2015; Frenken et al. 2017).

As different studies have shown (Scherngell and Barber, 2011; Scherngell and Lata, 2013; Wanzenböck et al., 2014; Enger, 2020), the collaboration in research projects could be explained by different factors, either **endogenous** (relating to the network structure) and **exogenous** to the extant network (relating to the node attributes).

Problem formulation

G(N;L): undirected graph, with N nodes and L links between nodes. We seek to predict what links are likely to be created in a future time *t'* given info at *t*



1. We compare link prediction accuracy from many different ML algorithms
2. Two models considered: with *all features* and with only the *exogenous* ones
3. Feature-importance by Average Partial Effects (APE)

Methodology and data

1. We combined three **Risis datasets**, the **EUPRO** dataset (a dataset providing information on R&D projects, participants and resulting networks of the EU FPs), the **RISIS-ETER** (database on European Higher Education Institutions) and **CWTS Publication** (a full copy of Web of Science)

2. We compare the performance of the proposed learning algorithms and compute the **prediction accuracy**, one embedding both **exogenous** and **endogenous** features, and one considering only exogenous features. We thus calculate the accuracy gap

3. For each learner, we then estimate the **average partial effects (APE)** function

4. We aggregate all the derivatives obtained in the previous step by averaging over them, thus obtaining a super learning derivative estimate (**elasticities**)

5. We also calculate **elasticities** to assess the percentage change of link probability induced by a given percentage change in the considered feature.

Endogenous and exogenous attributes

- **Betweenness centrality**, referring to the frequency that a university acts as a connection between a pair of other universities
- **Jaccard coefficient**, defined as the proportion of common neighbours in the total number of neighbours
- **Regional gross domestic product** (PPS per inhabitant), measured at regional level (source: EUROSTAT)
- **Core funding**, indicating the overall government funding available for a university (source: RISIS-ETER)
- **The average number of citations**, the average number of citations of the publications of a university, normalized for field and publication year (source: CWTS-Publication)
- **Number of students by ERC domain**, considered as a proxy of university size rescaled within the three ERC domains (source: RISIS-ETER)

ML methodology

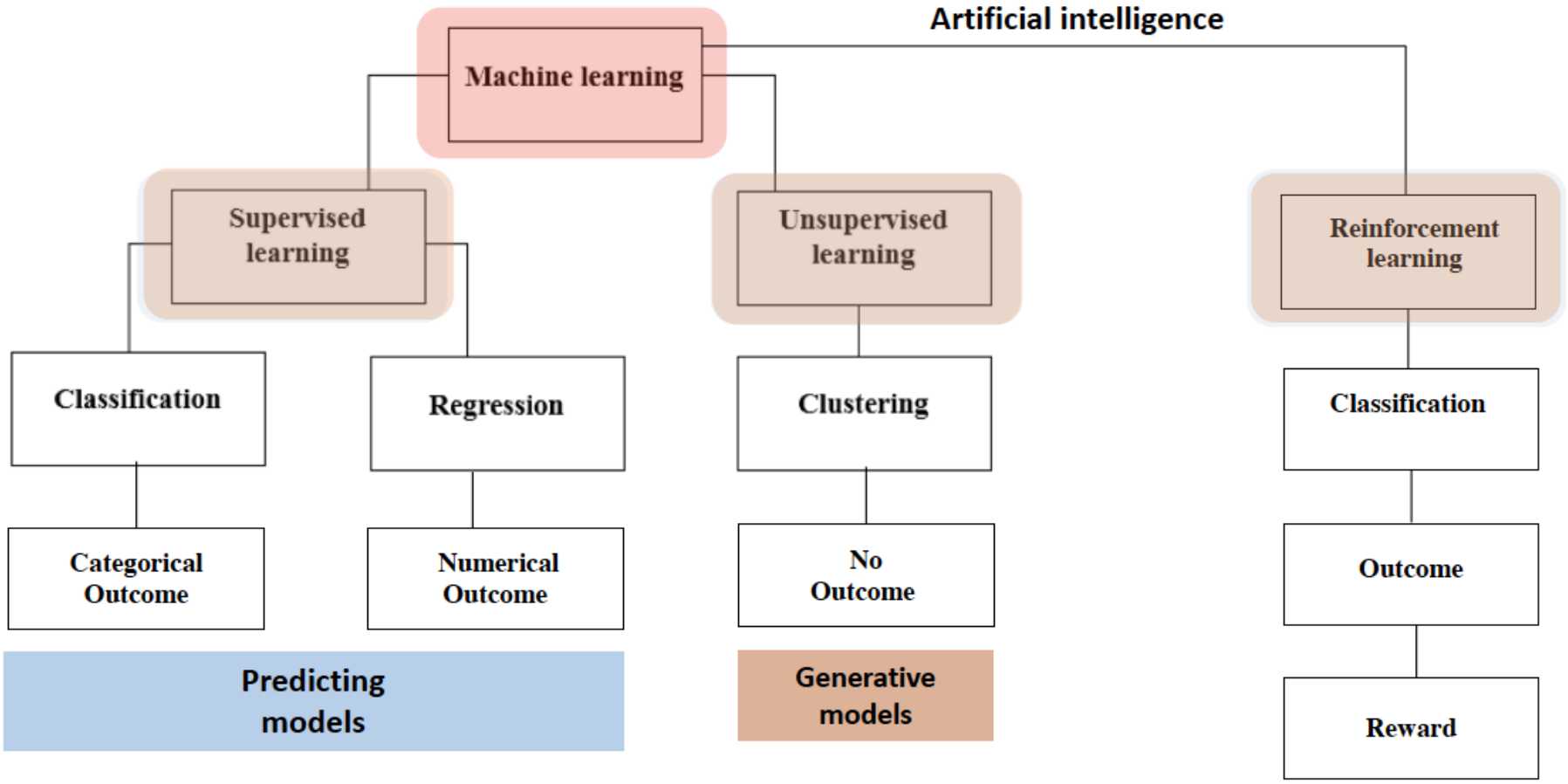
Machine Learning

A relatively new approach to **data analytics**, which places itself in the intersection between **statistics**, **computer science**, and **artificial intelligence**

ML objective

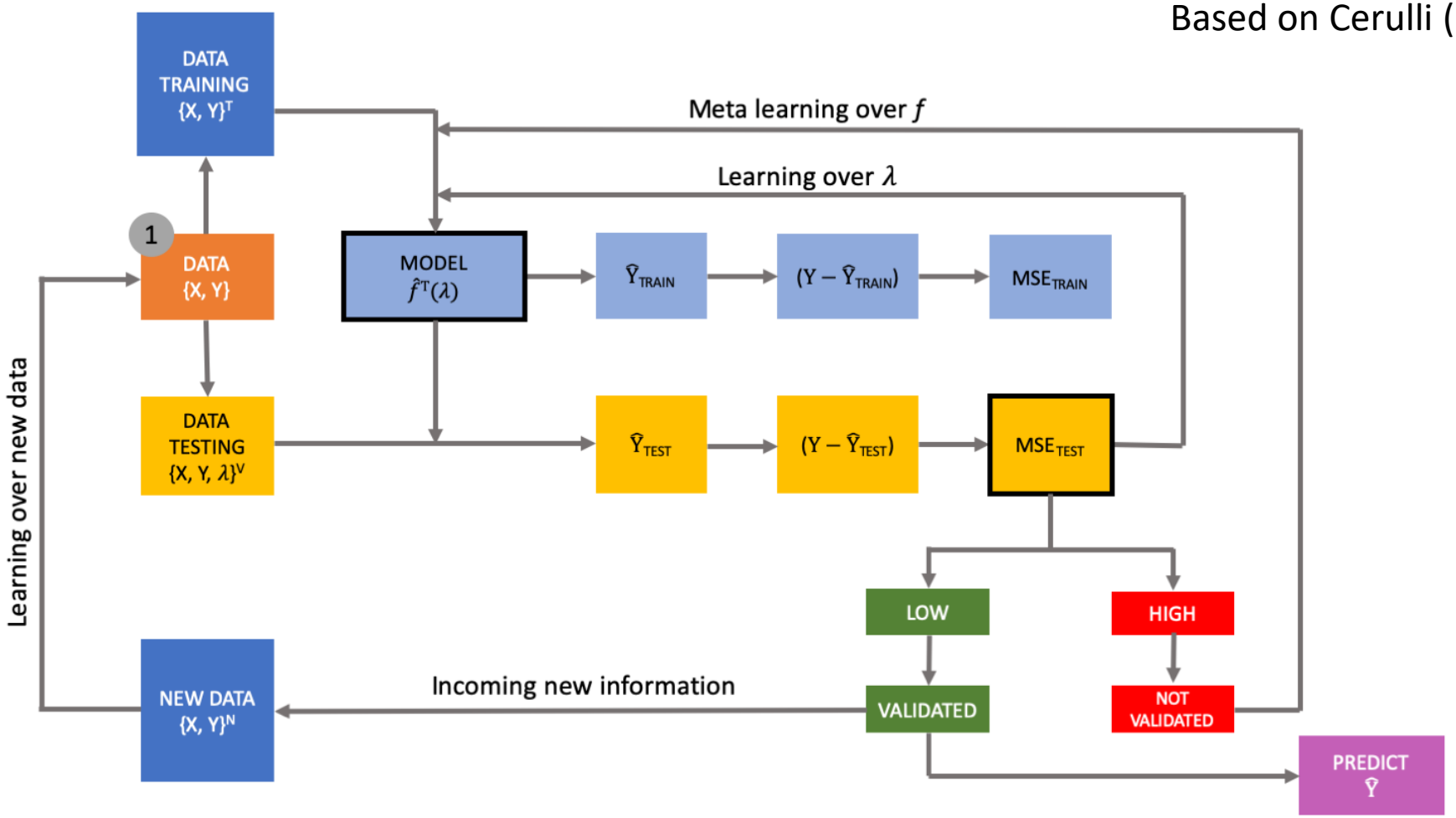
Turning **information** into **knowledge** and **value** by "**letting the data speak**"

Supervised and unsupervised learning

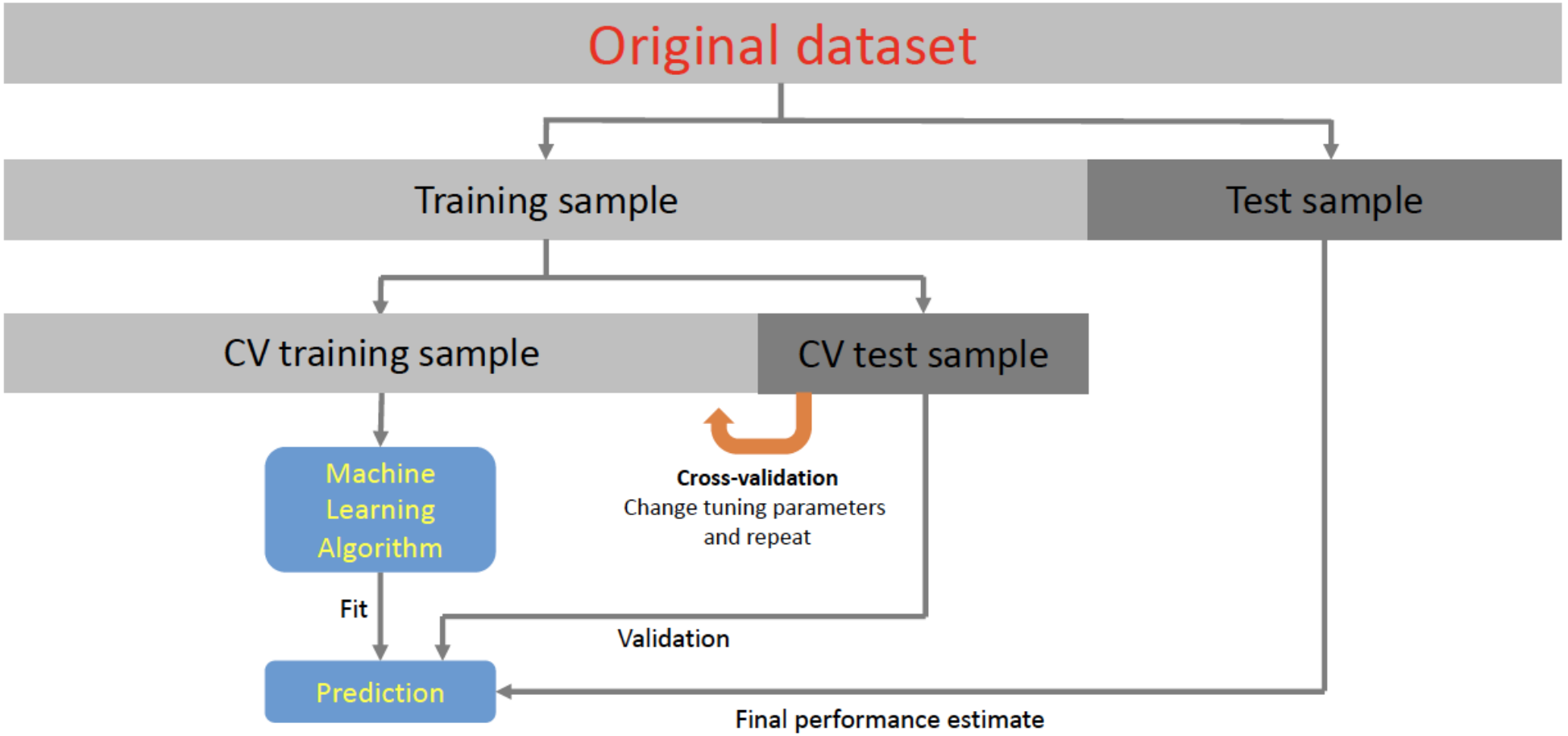


Our learning architecture

Based on Cerulli (2021)



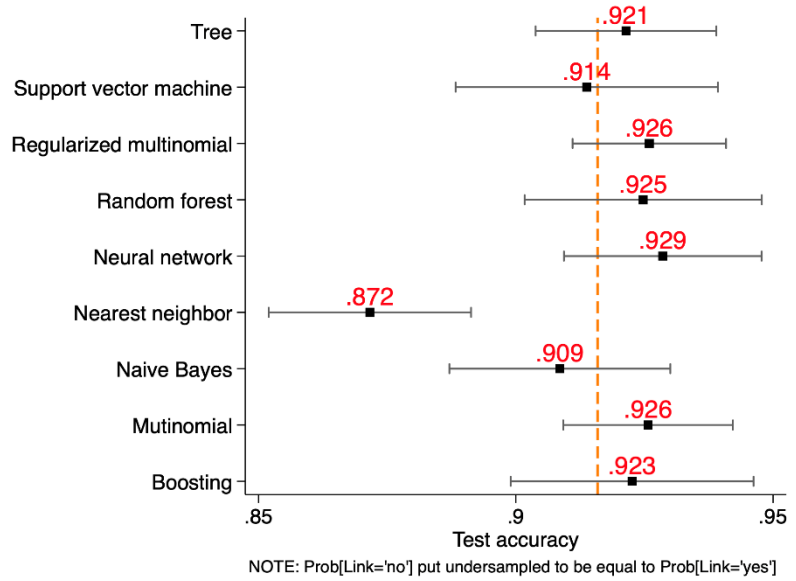
Model optimal tuning for prediction



Main results

Accuracy for SSH

Full model



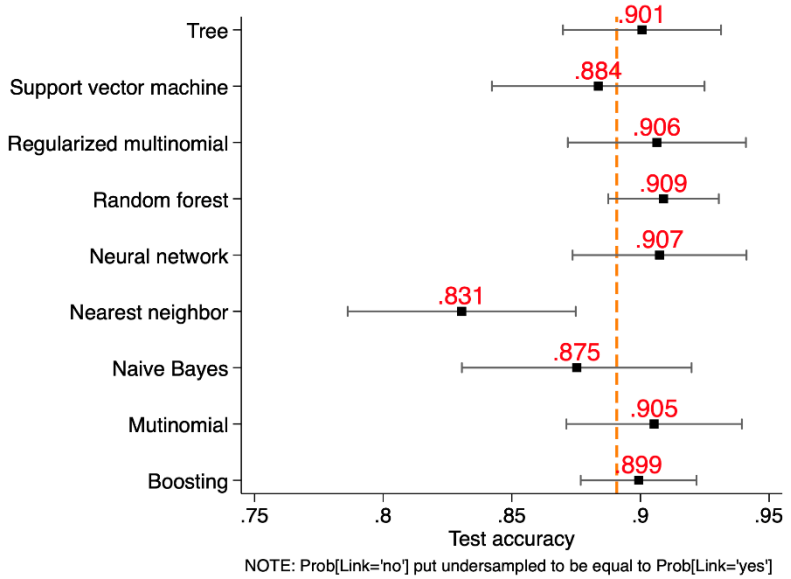
The knowledge of **endogenous** (i.e., network) attributes increases link prediction by around **30 points**

	TRAIN ACCURACY	TEST ACCURACY
Tree	.5891349	.5799
Support vector machine	.6579343	.5929983
Regularized multinomial	.6179249	.5965706
Random forest	.6270541	.5977614
Neural network	.6262866	.6013337
Nearest neighbor	.6682279	.5789474
Naive Bayes	.6238785	.6046678
Mutinomial	.6186658	.5941891
Boosting	.6146703	.5932365

Exogenous model

Accuracy for PE

Full model



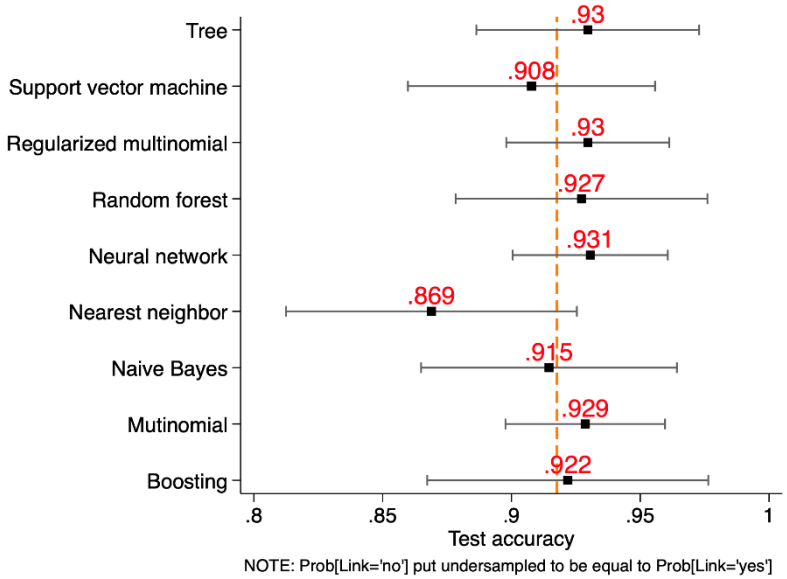
The knowledge of **endogenous** (i.e., network) attributes increases link prediction by around **20 points**

	TRAIN ACCURACY	TEST ACCURACY
Tree	.7422327	.7075595
Support vector machine	.7887643	.7057285
Regularized multinomial	.7269526	.7148836
Random forest	.7314213	.7282239
Neural network	.7348219	.7198535
Nearest neighbor	1	.7104368
Naive Bayes	.7197666	.7133142
Mutinomial	.7269744	.7148836
Boosting	.7335428	.7219461

Exogenous model

Accuracy for LS

Full model

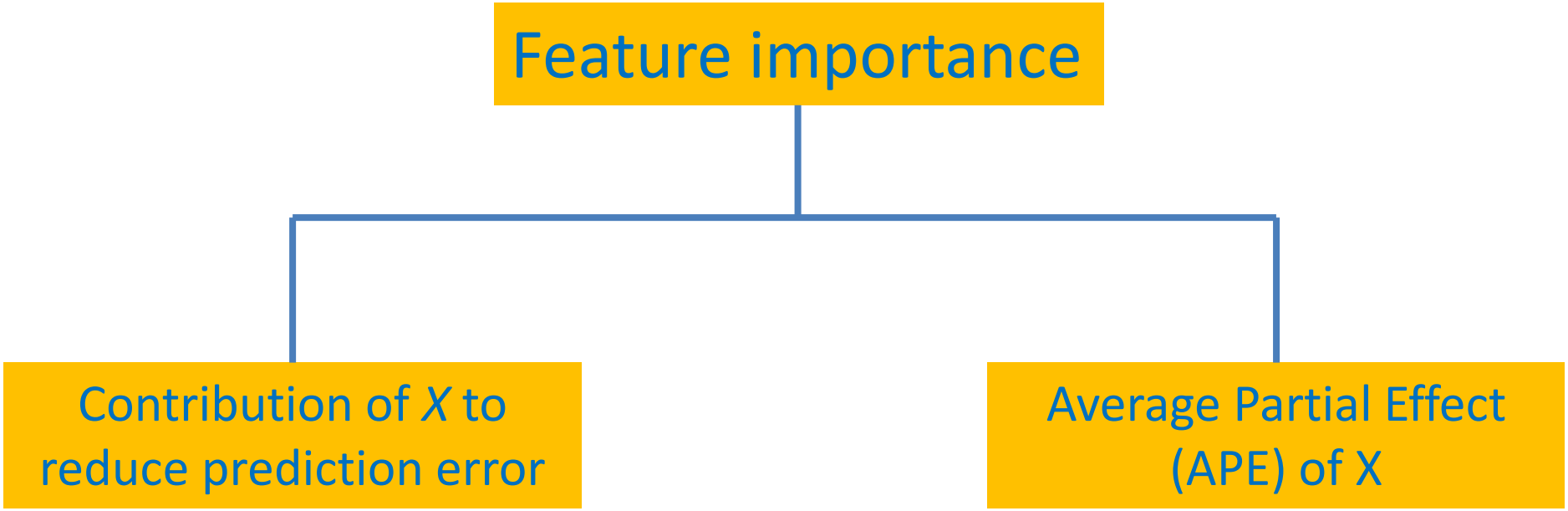


	TRAIN ACCURACY	TEST ACCURACY
Tree	.6873787	.6873786
Support vector machine	.7159106	.6800971
Regularized multinomial	.6966016	.6873786
Random forest	.6873787	.6873786
Neural network	.6881788	.6830097
Nearest neighbor	.722761	.6626214
Naive Bayes	.6717967	.6432039
Mutinomial	.6990829	.6810679
Boosting	.6873787	.6873786

The knowledge of **endogenous** (i.e., network) attributes increases link prediction by around **20 points**

Exogenous model

Measuring **feature importance** in ML



Does not provide effect's size and direction

Provide a effect's size and direction

We use this one !

Link probability's Average Partial Effect (APE)

$$\mathrm{APE}(y, x_j) = \frac{\partial E(y|x_j, \bar{X}_{-j})}{\partial x_j} = \frac{\partial Prob(y = 1|x_j, \bar{X}_{-j})}{\partial x_j}$$

Increment/decrement of the link probability for an infinitesimal change of the feature *x*, (all the other features held constant)

From derivatives to elasticities

APE is the link probability derivative at each point of the support of *x*

APE measures the shape of the relationship between link probability and the feature, but interpretation is tricky ("infinitesimal change")

ELASTICITY allows to measure the percentage change for the link probability for a 1 percent change in the feature

From derivatives to elasticities

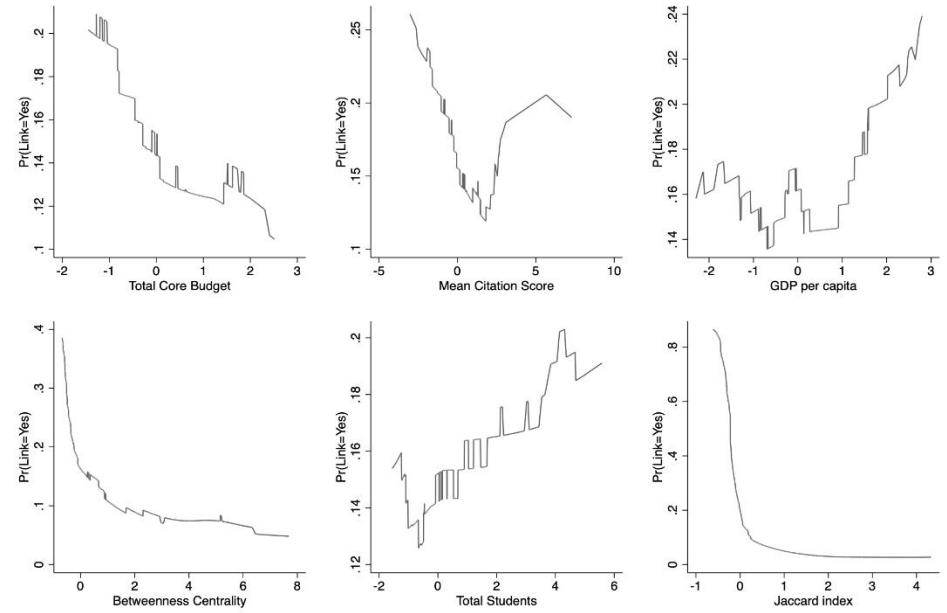
ELASTICITY: *percentage* change for the link probability for a *1 percent* change in the feature

$$\text{Elasticity} = \frac{\partial Prob(y = 1|x_j, \bar{X}_j)}{Prob(y = 1|x_j, \bar{X}_j)} \Big/ \frac{\partial x_j}{x_j}$$

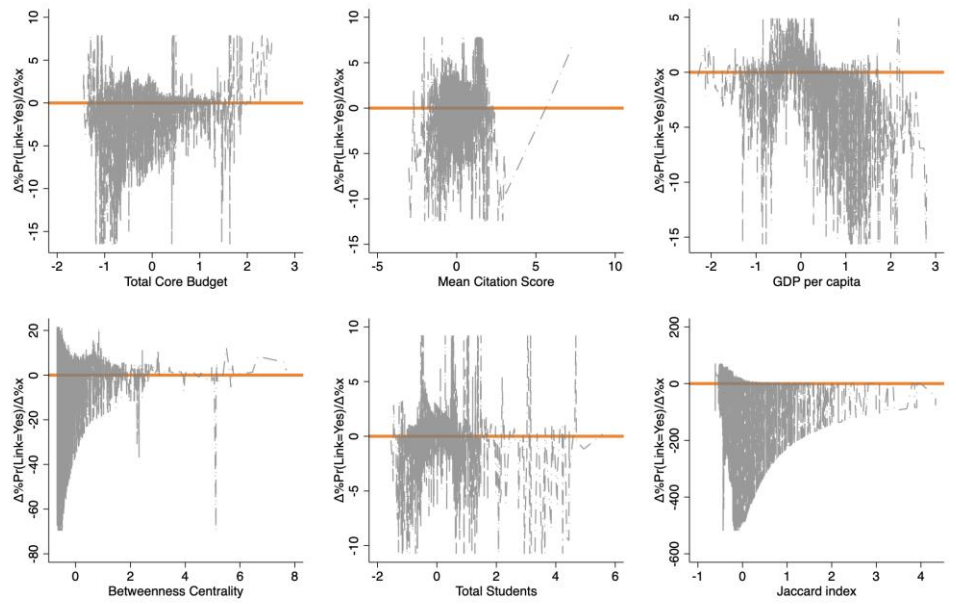
Link probability pattern by feature for SSH

$$Prob(y = 1|x_j, \bar{X}_{-j})$$

$$\frac{\partial Prob(y = 1|x_j, \bar{X}_j)}{Prob(y = 1|x_j, \bar{X}_j)} \Big/ \frac{\partial x_j}{x_j}$$



MEASURE: Marginal effect
METHOD: Average over all the learners

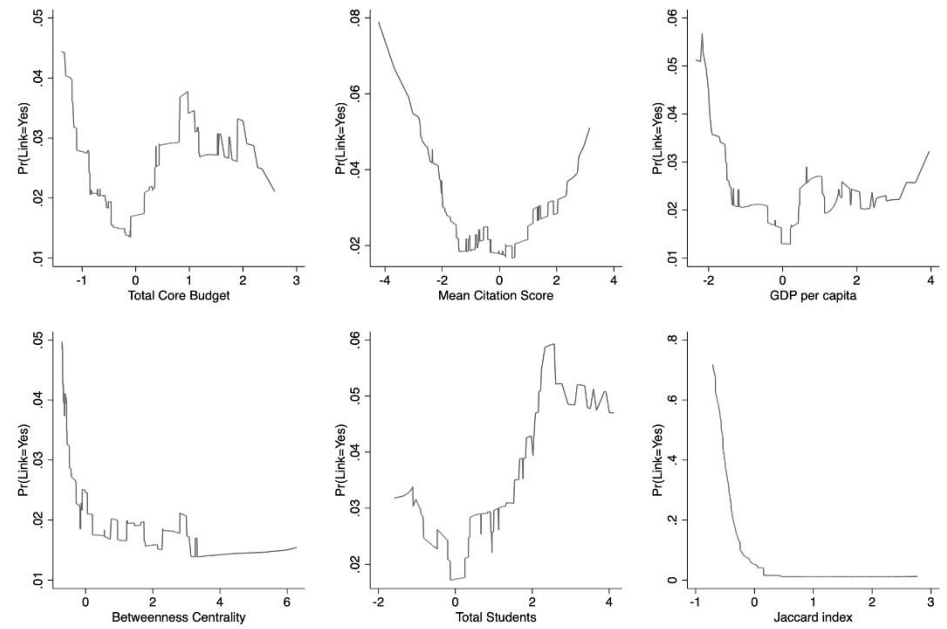


MEASURE: Elasticity
METHOD: Average over all the learners
ROBUSTNESS: Winsorized values
SCALE: % increase in the probabiity to link for a 100% increase in the feature

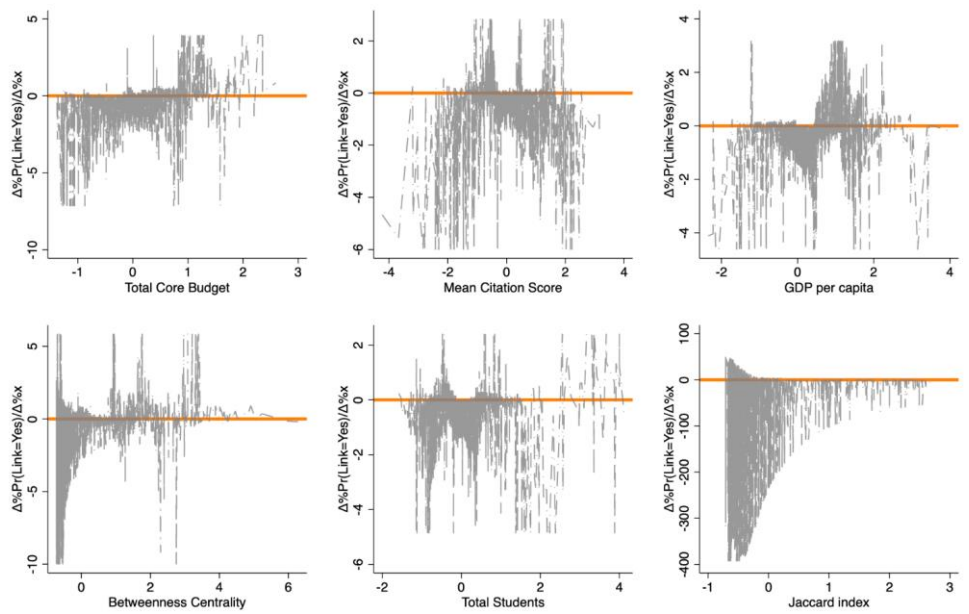
Link probability pattern by feature for PE

$$Prob(y = 1|x_j, \bar{X}_{-j})$$

$$\frac{\partial Prob(y = 1|x_j, \bar{X}_j)}{Prob(y = 1|x_j, \bar{X}_j)} \Big/ \frac{\partial x_j}{x_j}$$



MEASURE: Marginal effect
METHOD: Average over all the learners

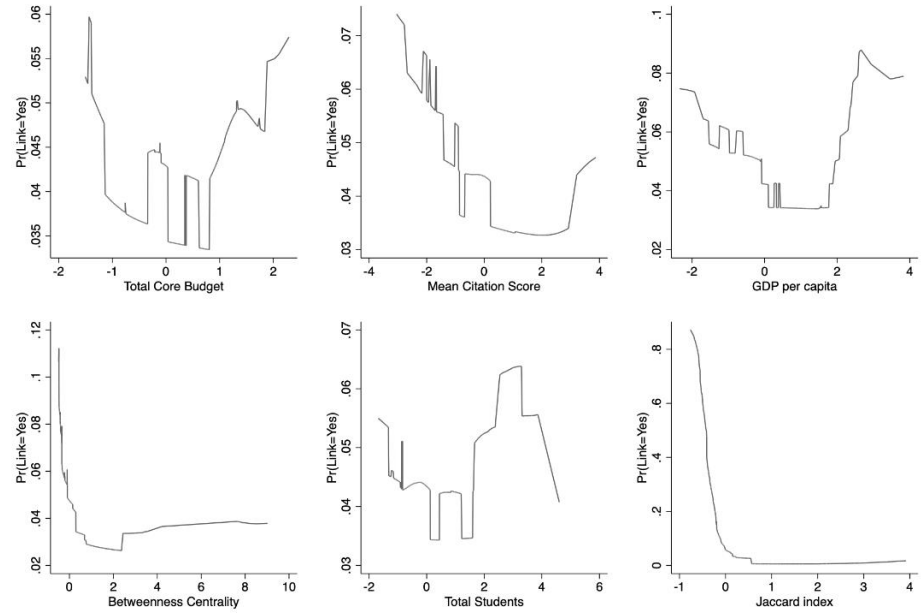


MEASURE: Elasticity
METHOD: Average over all the learners
ROBUSTNESS: Winsorized values
SCALE: % increase in the probabiity to link for a 100% increase in the feature

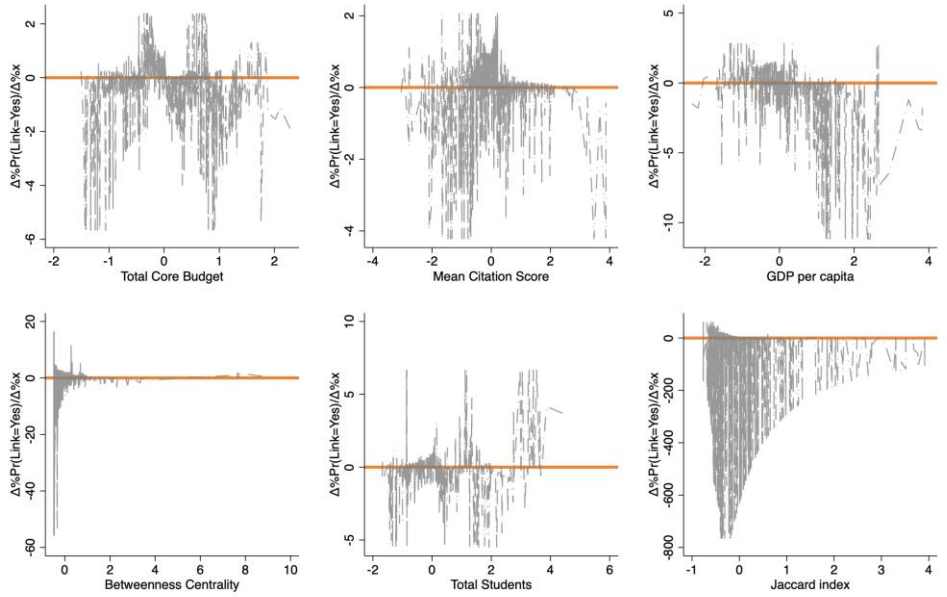
Link probability pattern by feature for LS

$$Prob(y = 1|x_j, \bar{X}_{-j})$$

$$\frac{\partial Prob(y = 1|x_j, \bar{X}_j)}{Prob(y = 1|x_j, \bar{X}_j)} \Big/ \frac{\partial x_j}{x_j}$$



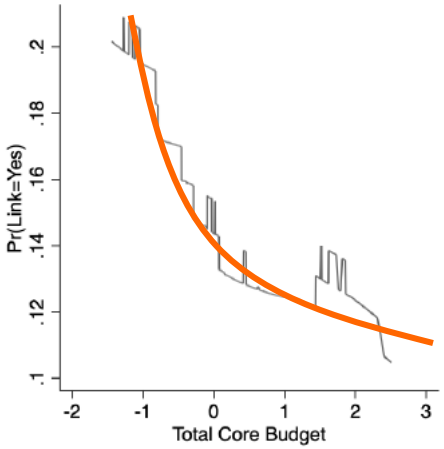
MEASURE: Marginal effect
METHOD: Average over all the learners



MEASURE: Elasticity
METHOD: Average over all the learners
ROBUSTNESS: Winsorized values
SCALE: % increase in the probabiity to link for a 100% increase in the feature

Core Funding exhibits different patterns in the three domains

SSH

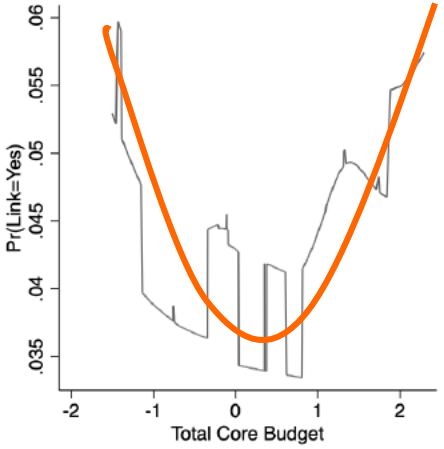


The larger the Core Funding, the smaller the probability of a pair to get linked

↓

Budget constrain effect

LS

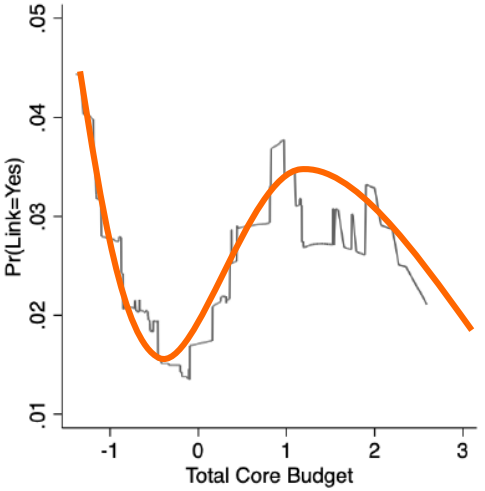


From a certain threshold of Core Funding on, collaborating becomes likelier

↓

Scale complementarities

PE



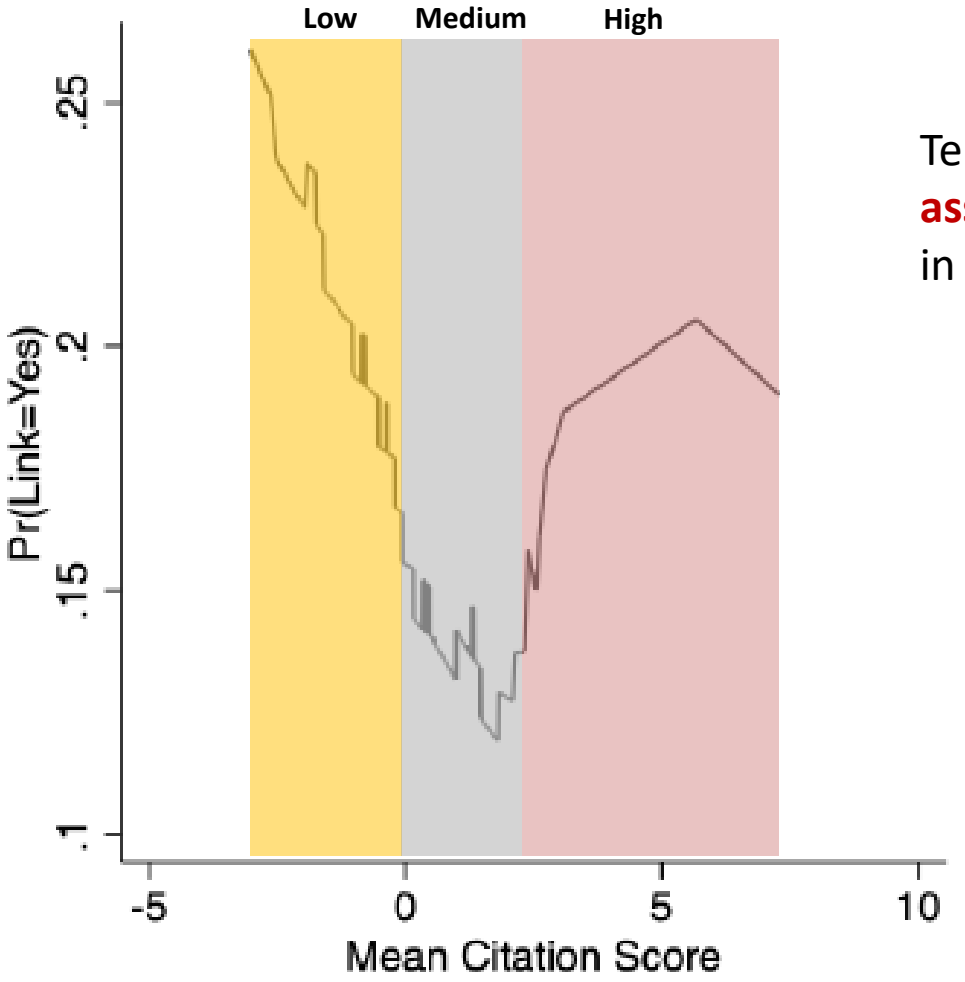
Larger scale requires collaboration but less than in LS

↓

Infrastructure effect

Mean Citation Score exhibits U-shaped pattern

Pairs characterized either by low or high MCS tend to link together

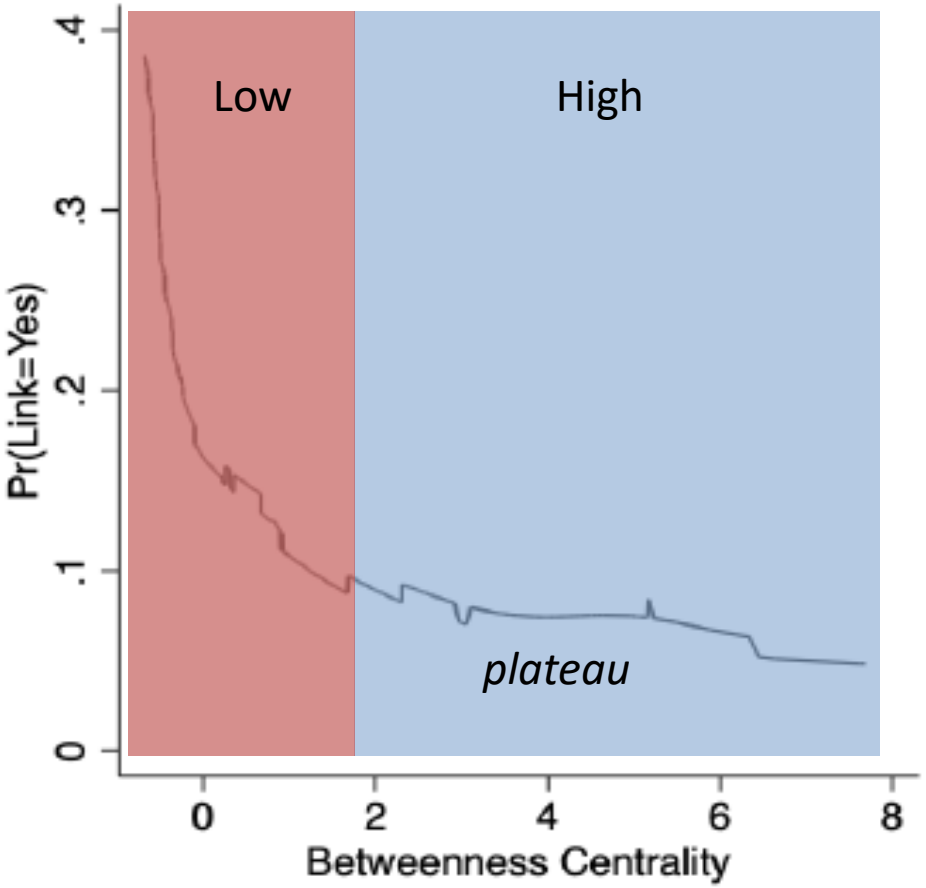


Tendency of universities to **associate** with universities **similar** in terms of **output quality**

Homophily

Betweenness Centrality exhibits a decreasing pattern

Poorly central pairs in terms of betweenness tend to link more than more central pairs

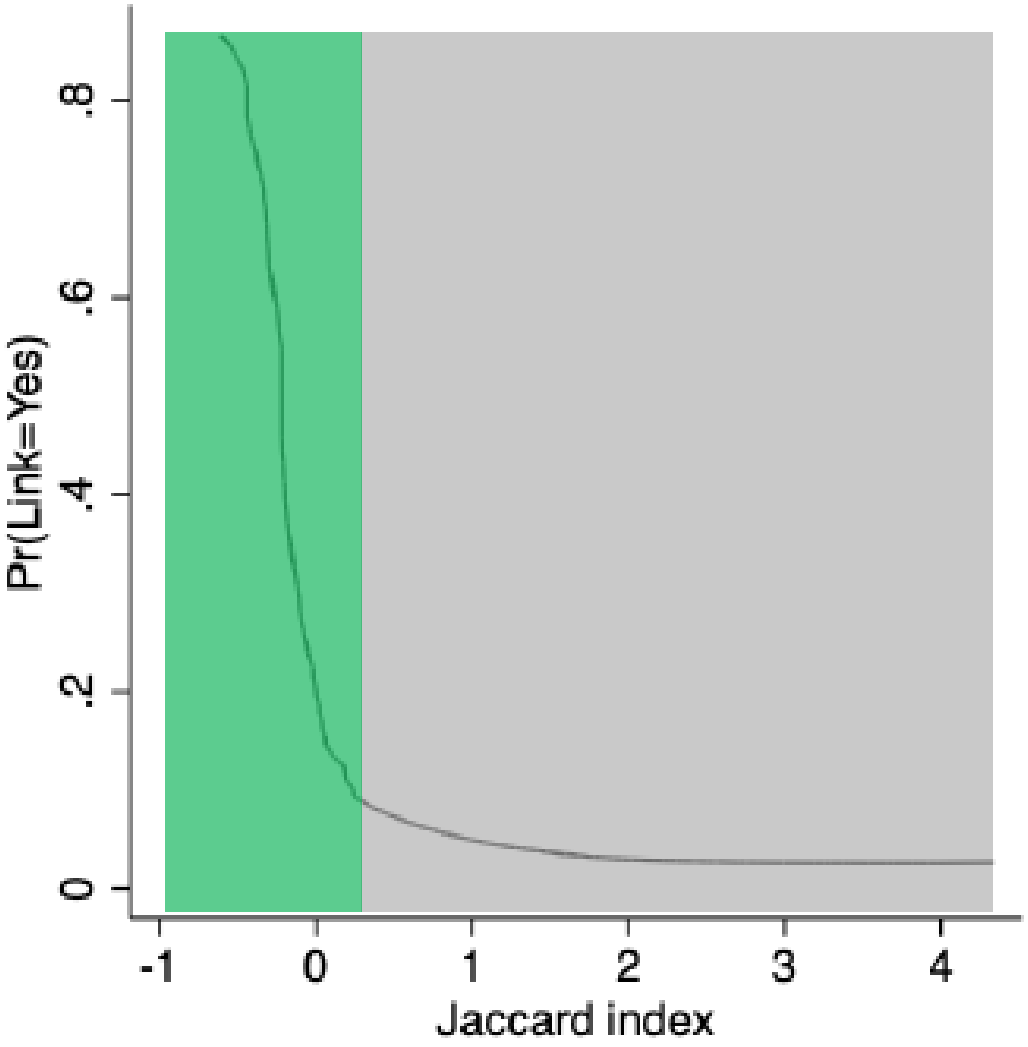


If two universities have high centrality in the network, they tend not to link in a direct way. They represent the center of two disconnected ***archipelagos*** (*communities*)

Cognitive dissimilarity

Jaccard Similarity Index exhibits a decreasing pattern

More similar nodes do not tend to collaborate (when the % of common neighbors is high)



From a **resource-based viewpoint**, two universities with similar **knowledge-base** are poorly attracted, as they look for **complements**, not **substitutes**

Cognitive complementarity

Conclusions

- Link prediction accuracy larger than 90% for pretty all the machine learning methods
- By removing endogenous features, prediction accuracy drops down in all domains by a 25 points on average
- *Jaccard index* and *Betweenness Centrality* important to predicting links in all domains
- *Jaccard index*, *Betweenness Centrality* and *Mean Citations Score* exhibit very stable patterns in all domain. *GDP per-capita* shows a less strong similar pattern
- In the SSH domain, Core funding plays a different role than in PE and LS