Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns

Soumya Banerjee

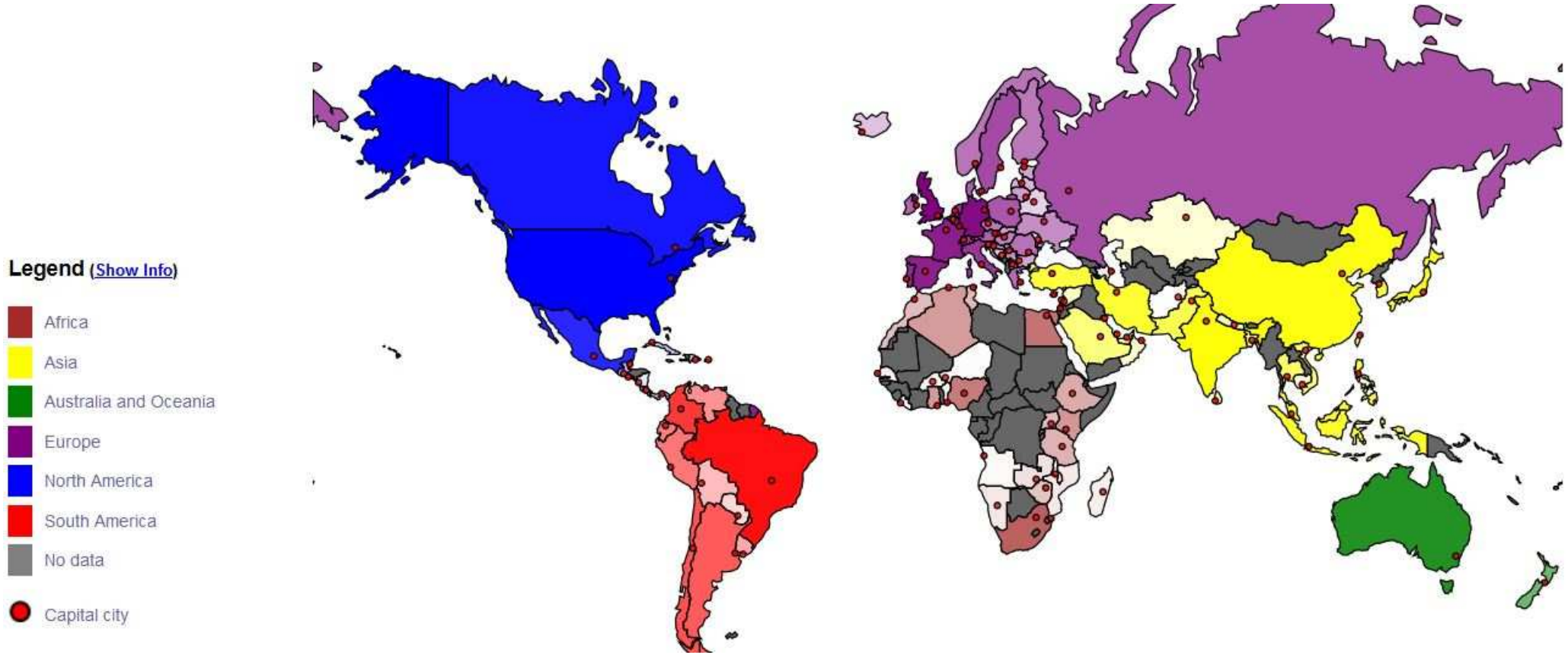
Motivation

- Scientific collaboration networks are an important component of scientific output
- Contribute significantly to expanding our knowledge and to the economy and gross domestic product of nations
- We examine a dataset from the Mendeley scientific collaboration network.
- We analyze this data using a combination of machine learning techniques and dynamical models.

Motivation

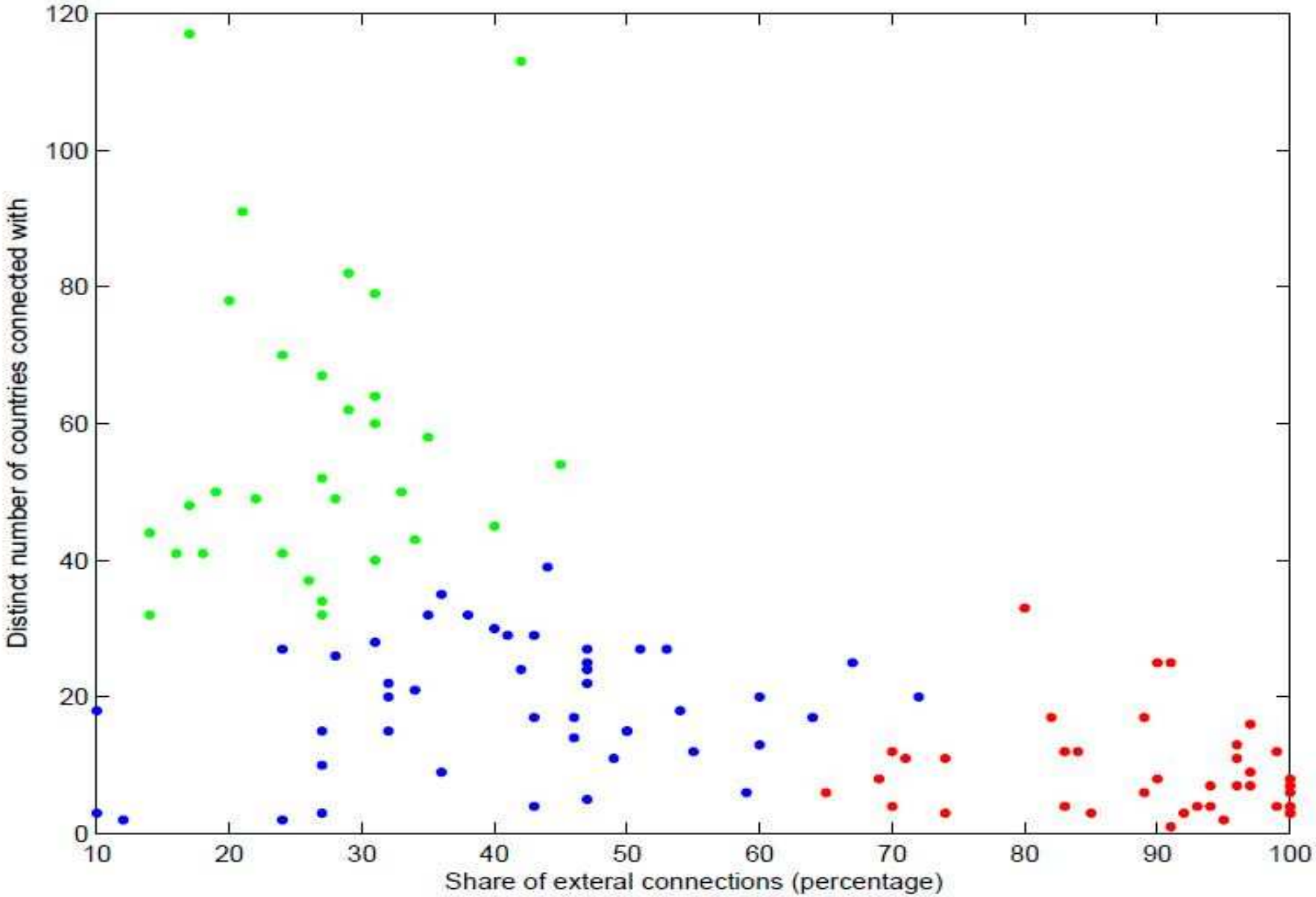
- We find interesting clusters of countries with different characteristics of collaboration. Some of these clusters are dominated by developed countries that have higher number of self connections compared with connections to other countries.
- Another cluster is dominated by impoverished nations that have mostly connections and collaborations with other countries but fewer self connections.
- We propose a complex systems dynamical model that explains these characteristics.
- Our model explains how the scientific collaboration networks of impoverished and developing nations change over time.
- We find interesting patterns in the behaviour of countries that may reflect past foreign policies and contemporary geopolitics.

Patterns in Global Collaboration Data



http://labs.mendeley.com/collab-map/

Patterns in Global Collaboration Data



Patterns in Global Collaboration Data

- In order to explain these patterns, we propose a dynamical model for how the links are actually formed
- Developing countries might preferentially seek out collaborations with richer countries or richer countries might help disadvantaged nations (e.g. Liberia has 100% external connections).
- We simulate the time evolution of connections. Our dynamical model correctly predicts the broad observations of scientific connections of countries.

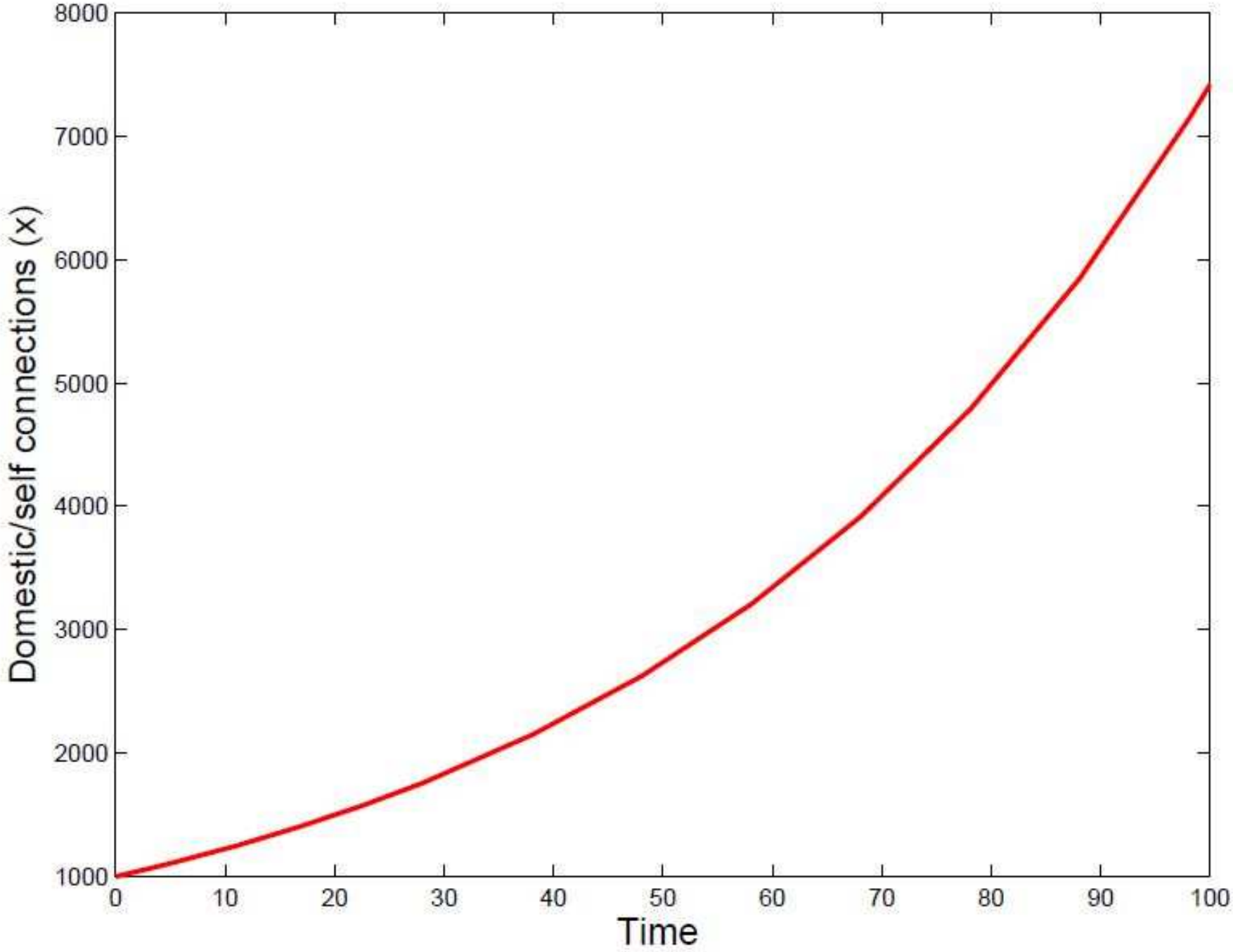
Dynamical Model

Our dynamical model has two compartments: developing countries (x) and developed countries (y). Developing countries grow their scientific expertise and self-connections at a rate proportional to $\alpha * x$ (self-growth) and also by interacting with developed nations at a rate proportional to $\beta * x * y$. Developed nations (y) are assumed to have reached equilibrium of growth. The model, represented as differential equations is shown below:

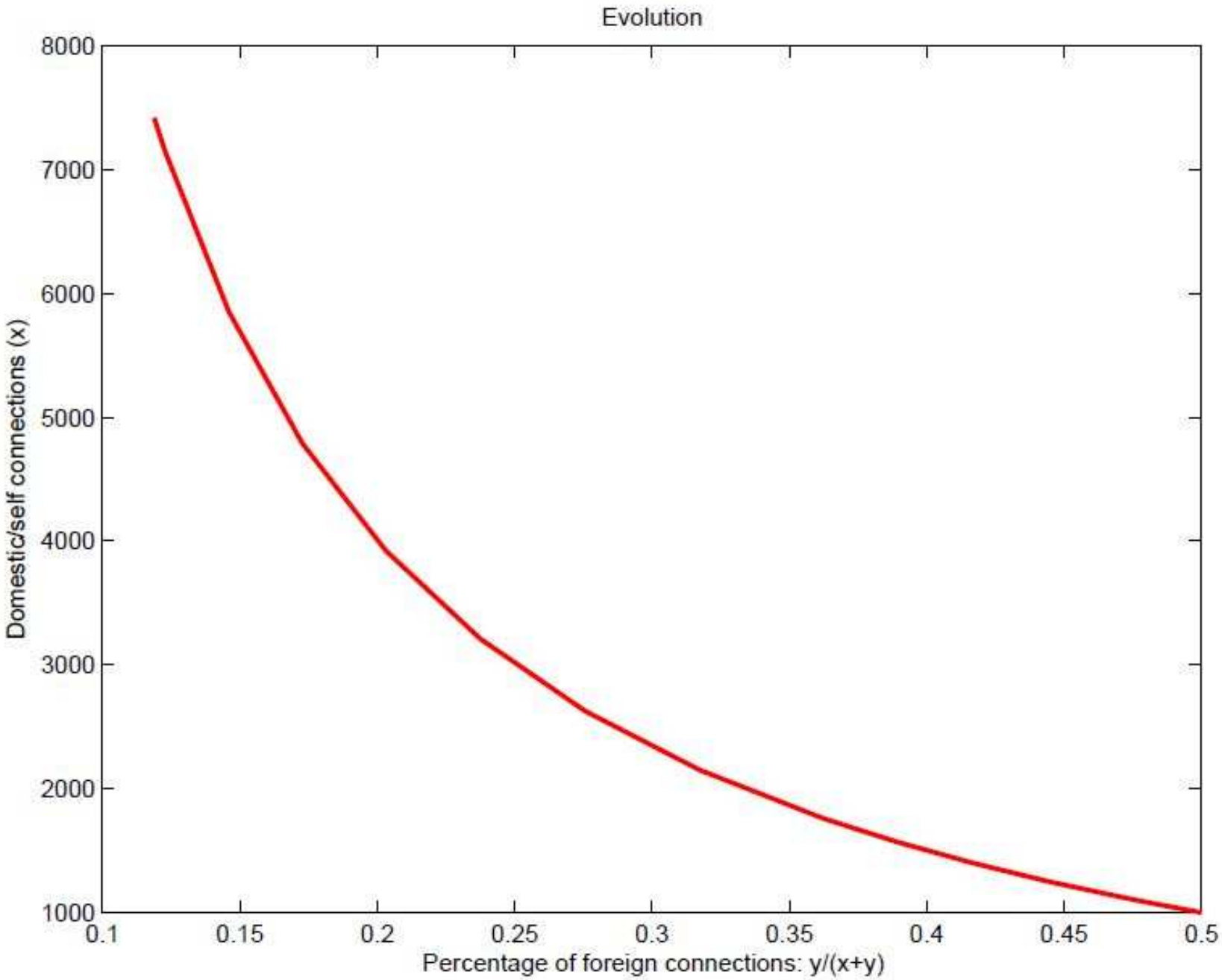
$$\frac{dx}{dt} = \alpha x + \beta xy$$

$$\frac{dy}{dt} = 0$$

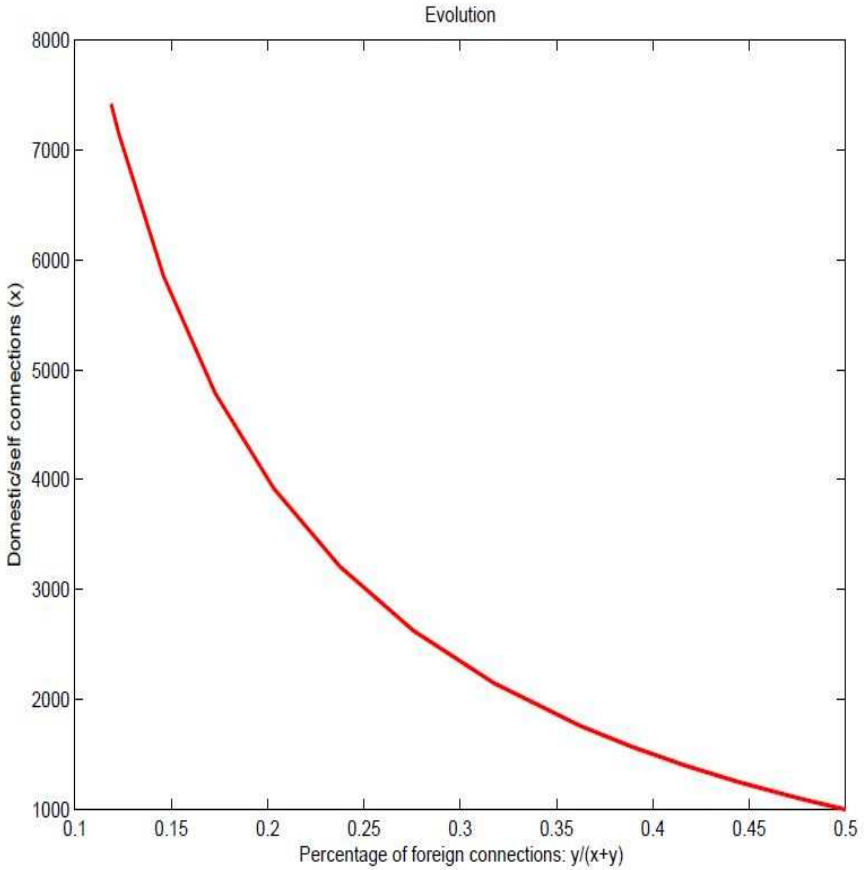
Model Simulations

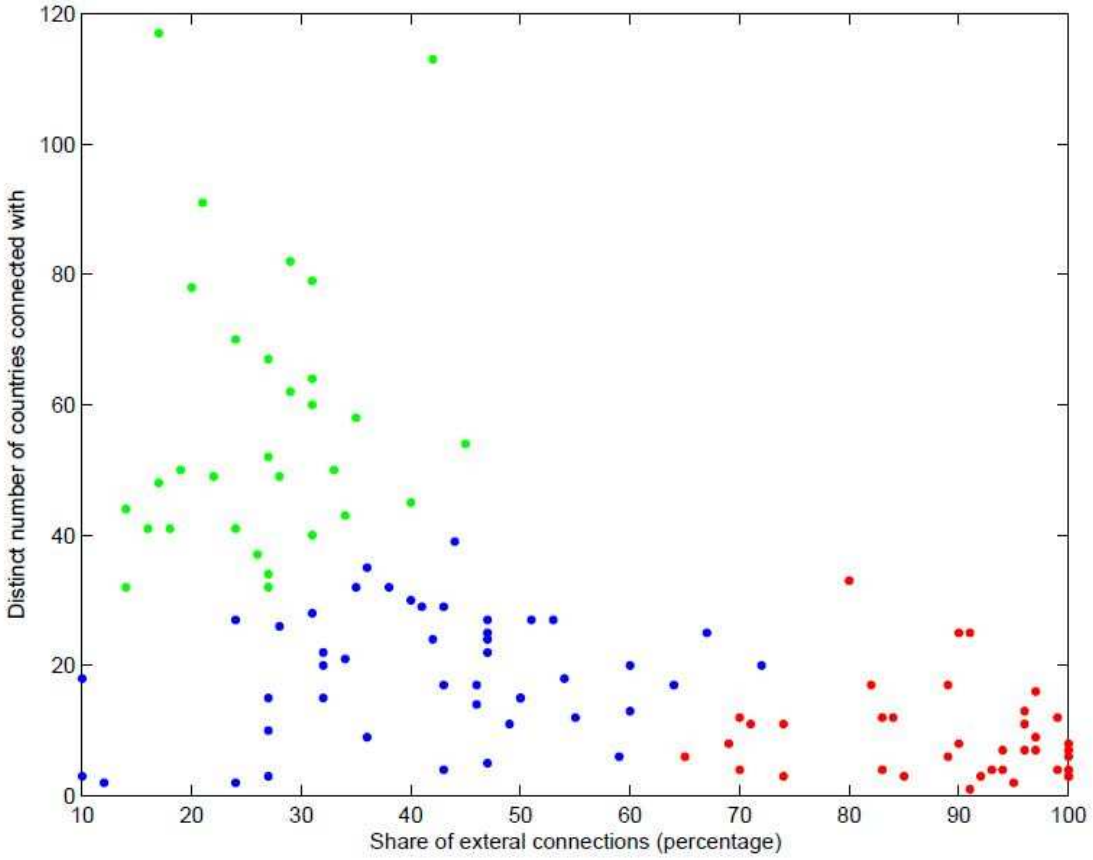


Model Simulations



Model Simulations vs. Actual Data





Empirical Data

Country Name	share_of_external_connections	distinct_no_of_countries_country_is_connected_with
United States	17	117
United Kingdom	42	113
Germany	21	91
France	29	82
Netherlands	31	79
Canada	20	78
Australia	24	70
Spain	27	67
Italy	31	64
Sweden	29	62
Switzerland	31	60
Belgium	35	58
India	45	54
Austria	27	52
Portugal	19	50
Argentina	33	50
South Africa	22	49
Denmark	28	49
Brazil	17	48

Observations

- The UK has a very high percentage of foreign connections; this maybe because of its colonial past.
- Less developed or poorer countries usually have a very high percentage of foreign connections.
- Iran has a large number of foreign connections; it is interesting that this is so despite foreign sanctions imposed against it.
- The richest countries have more distinct foreign connections.
  - may suggest that the rest of the world wants to collaborate and form connections with others
  - could also be driven by interest in issues relevant to poorer countries (like tropical diseases, socio-economic research into poverty and archaeological research in countries in Africa)

Observations

- India has more foreign connections than China.
- The US has a lower percentage of foreign connections but in absolute numbers it has the highest number of connections.
- The highest percentage of foreign connections is usually occupied by very poor countries (presumably they are trying to build capability in science and technology by collaboration), e.g. Liberia has 100% foreign connections.
- Some countries like El Salvador have a low percentage of foreign connections (this could be as a result of the fact that the country was embroiled in civil war for a long time). The policy implications are that it shows a need for targeted relief at starting active science and research programs in these nations.

Observations

- South Korea, Japan, Taiwan and Germany have a low percentage of foreign connections (they have a large number of absolute connections and it is possible that they have invested very heavily in their own science programs).

- Cuba has very few connections (this is again possibly due to sanctions imposed against it when the data was collected).
  - Sanctions have now been lifted and it would be interesting to observe how that affects science and development

Conclusions

- We find interesting patterns and clusters of countries with different characteristics of collaboration.
- We use machine learning techniques and dynamical models to analyze this data
- Our model and analysis gives insights and guidelines into how scientific development of developing countries can be guided.
- This is intimately related to fostering economic development of impoverished nations and creating a richer and more prosperous society.

Outreach

- Play with dynamical models, code and data
- All resources will be available at
  - https://bitbucket.org/neelsoumya/public_open_source_datascience

Acknowledgements

- Mendeley Labs – for sharing the data
- Three referees for great suggestions

Data

- Mendeley collaboration data (from http://labs.mendeley.com/collab-map/)
- Processing on data
  - connection between two researchers as co-membership in a Mendeley group.
- Only private groups are considered
- Only groups created or joined before March 2013 are considered