AFRL-AFOSR-VA-TR-2019-0264

The fundamentals of predictability of scientific success

Albert-Laszlo Barabasi
NORTHEASTERN UNIVERSITY

07/19/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTB1
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10-09-2019	Final Performance	01 Feb 2015 to 31 Jan 2019

4. TITLE AND SUBTITLE
The fundamentals of predictability of scientific success

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-15-1-0077

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Albert-Laszlo Barabasi

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NORTHEASTERN UNIVERSITY
360 HUNTINGTON AVE
BOSTON, MA 02115 US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTB1

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2019-0264

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
During 2018, our team has been focusing on understanding scientific success in a couple of aspects, including scientific ability, gender, social network, team, country, discipline, move, etc. To be specific, as we discuss next, we tackled several key questions, aiming to propose quantitative and predictive models to understand, explain and predict scientific success.
1. [Ongoing] Data curation and disambiguation: This is one of the most critical tasks that facilitates the implementation of the whole line of work, by providing more accurate and expansive datasets. Tackling the author name ambiguity problem, we developed a method to identify authors with high accuracy, collaborating with a group from University of Massachusetts Amherst. Specifically, we employed several layers of information, from an author's affiliation to email address and the list of his/her collaborators, allowing us to uniquely identify each scientist. This method outperforms previous efforts in disambiguating the Web of Science dataset by providing a higher accuracy and covering a broader extent of data.
2. [Ongoing] Impact of network on scientific success: We have started a new line of research to separate the social impacts of the scientific community on the perceived success of a scientist. Our analysis assumes that the impact of a scientist is strongly determined by the structure of her professional interaction network, defined as the list of scientists with whom one has collaborated or worked at the same institution or got to know them through a common collaborator. During last year our effort has focused on extracting the pertinent interaction network of scientists, which provides the primary information to understand the effect of the interaction network on scientific success.
3. [Published] Evolution of impact and productivity in scientific careers. In a paper published in Science, we quan

15. SUBJECT TERMS
scientific success, scientific excellence, citation patterns, science emergence, paper citations

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			PARRA, ENRIQUE
Unclassified	Unclassified	Unclassified	UU		19b. TELEPHONE NUMBER *(Include area code)* 703-696-8571

AFOSR – Report Submission Form

Air Force Office of Science and Research
875 Randolph Street
Suite 325 Room 3112
Arlington, VA 22203

1. Report Type
Annual Report

Primary Contact E-mail

c.mannett @northeastern.edu

Primary Contact Phone Number

617-373-6869

Organization / Institution name

Northeastern University

Award Information

Grant/Contract Title
The full title of the funded effort.

The fundamentals of predictability of scientific success

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-15-1-0077 FA9550-XX-X-XXXX

Principal Investigator Name
The full name of the principal investigator on the grant or contract.

Albert-László Barabási

Program Manager
The AFOSR Program Manager currently assigned to the award

Enrique Parra

Report Information - Annual Report

For an annual report, the reporting period start date is either start date of the grant, if this is the first report, or 1 day after the due date of the previous report. The end date is due date of this report.

Reporting Period Start Date

1 February 2018

Reporting Period End Date

31 January 2019

Report Abstract:

In the Abstract section, please list any accomplishments that have been made since the last report submission (or since the beginning of the award if this is the first report). Please **do not** type "see report" here, include at least an abstract, **250 words or more**, of the accomplishments mentioned in your report.

During 2018, our team has been focusing on understanding scientific success in a couple of aspects, including scientific ability, gender, social network, team, country, discipline, move, etc. To be specific, as we discuss next, we tackled several key questions, aiming to propose quantitative and predictive models to understand, explain and predict scientific success.

1. [Ongoing] **Data curation and disambiguation**: This is one of the most critical tasks that facilitates the implementation of the whole line of work, by providing more accurate and expansive datasets. Tackling the author name ambiguity problem, we developed a method to identify authors with high accuracy, collaborating with a group from University of Massachusetts Amherst. Specifically, we employed several layers of information, from an author's affiliation to email address and the list of his/her collaborators, allowing us to uniquely identify each scientist. This method outperforms previous efforts in disambiguating the Web of Science dataset by providing a higher accuracy and covering a broader extent of data.

2. [Ongoing] **Impact of network on scientific success***:* We have started a new line of research to separate the social impacts of the scientific community on the perceived success of a scientist. Our analysis assumes that the impact of a scientist is strongly determined by the structure of her professional interaction network, defined as the list of scientists with whom one has collaborated or worked at the same institution or got to know them through a common collaborator**.** During

last year our effort has focused on extracting the pertinent interaction network of scientists, which provides the primary information to understand the effect of the interaction network on scientific success.

3. [Published] **Evolution of impact and productivity in scientific careers**. In a paper published in Science, we quantified the evolution of impact and productivity throughout scientific careers, to explore how impact and productivity change over a scientific career, and to model scientific careers in quantitative and predictive terms. With analysis on over 2,800 scientists with publication record spanning over 20 years, we find that the highest-impact work in a scientist's career is randomly distributed. This allows us to develop a quantitative model to systematically untangle the role of productivity and luck in each scientific career. The model assumes that each scientist selects a project with a random potential and improves on it with a factor Q, which make up the impact of its resulting publication. The factor Q characterize a scientist by capturing his/her ability to take advantage of the available knowledge in a way that enhances ($Q > 1$) or diminishes ($Q < 1$) the potential impact of a paper. A scientist's Q is independent of career stage, which is in contrast with all current metrics of excellence. Therefore Q can be estimated with early stage performance and used to predict independent recognitions like Nobel prizes.

4. [Ongoing] **Mapping out Network Science**. We started to explore success within network science itself, using the tools of network science, in hope of providing a clear picture of network science and reveal its structure and trend. Building on our work in defining physics papers, published last year, we developed a method to automatically detect network science papers, by propagating labels from a small set of "core" of manually collected network science papers on a citation network. This method is based on the hypothesis that a paper is considered a network science paper if it is consistently citing and cited by network science papers. We identified 40,000 network science papers from a total of 41 million in Web of Science, most of which are published in physics, math and life sciences journals. The papers make up a hub-and-spoke topology with a couple of high impact papers in center positions, while the highest impact network scientists show a clear division into two communities: one with focus on math and physics while the other with focus on life sciences.

5. [Ongoing] **Modeling the Gender Gap in Science**. We focus on the gender gap in academia, capturing the well-known fact that female scientists are underrepresented in many disciplines and countries. Although this is a well-documented, its driving force remains unclear. We analyzed the scientific careers of over 300,000 scientists from Web of Science, checking key features of male and female scientists including productivity, impact, and longevity. We noticed that the gender gap in longevity is strongly related with productivity, finding that female scientists produce less publications mostly because they are staying shorter within academia, against the stereotype that female scientists are less n productive. We find that controlling the excessive dropout rate of female scientists helps significantly reduces the gender gap in career performance.

6. [Ongoing] **Scientific boundary**. The legacy of history and geography forged a global scientific enterprise in which the scientific resources underneath are not evenly distributed, resulting in biased knowledge and talent flow, imbalanced research funds, inflated reputations. The large scale, and continuity of citation data make it possible to explore the citing behaviors of

scientists in diverse geographical regions. At first, we construct a region based knowledge flow network by parsing the geographical information related to each author's affiliation list(s) from the Web of Science dataset; Then we use a novel model to predict the patterns of collective citing behaviors and quantify the underlying factors (distance, border etc.) which influence the knowledge flow; Lastly, we apply this method to evaluating science, both for country and continent level, with reasonable expectations.

7. [Ongoing] **Understanding effort and success in teams**. A multitude of creative fields, from scientific discovery to software development, benefit from the creative output of coordinated groups of individuals. Here we explore the effect of team size in knowledge production: does the team size increase (social facilitation) or decrease (social loafing) the effort per member? In a quantitative Science of Science, large-scale online records are analyzed, from 6 million software development teams, to online collaboration in Wikipedia and a massive multiplayer online game. We find that the success of a team, as it grows, is increased in low-success projects while larger teams have less successful individuals. The goal of this work is to uncover general predictive mechanisms in team success, and check to which extent these mechanisms hold in co-author teams of scientific publications.

8. [Ongoing] **Interdisciplinarity in the Nobel Prize**. Despite the oft-cited importance of interdisciplinarity in science, there is no agreed definition in the Science of Science of what an interdisciplinary work is. In this project, we systematically explore a working quantitative definition of interdisciplinarity by considering statistical patterns in citation networks, and as an application reveal the role that interdisciplinarity plays in the Nobel Prize. With data collected from the Nobel Prize website, we analyze a large number of papers of physics, chemistry and the life sciences, exploring field-specific biases in the award of prizes, illuminating the mechanisms of scientific recognition. For example, given the amount of papers that are bridging physics and the life sciences, it is statistically less likely that a Nobel Prize winner is found in this interdisciplinary configuration. We are planning to explain the phenomenon in the future and hope to achieve a better understanding of the patterns of how high impact publications are related with interdisciplinary research.

9. [Manuscript under review] **Mapping the Knowledge Space**. We extended the analysis of the knowledge space along several different directions. First, we produced an analysis at a finer geographical resolution by looking at the knowledge production at the city level, rather than country level (as previously studied). Second, we refined the mapping of the knowledge space by using different “similarity measures” and by comparing the results obtained. Third, we looked at the temporal evolution of the knowledge topology. Fourth, we look at how different geographical entities (i.e. countries or cities) evolve their knowledge capacities over time given their past positioning in the knowledge space.

10. [Ongoing] **Mapping the Knowledge Space, Visualization Tool.** To improve our understanding of the dynamics studied at task 9 and to help illustrating our results to a more general public, we have started developing an online visualization tool that allows the user to explore the features of the knowledge space, given the individual the ability to change some of the assumption made (e.g. the concept of knowledge similarity) to observe what changes that causes in the knowledge mapping.

11. [Ongoing] **Knowledge Complexity and Country/City Fitness.** Using the data and some of the results described at task 9, we have started working on producing a metric that can help us understand the level of complexity embedded in some specific branches of the scientific knowledge as it emerges by looking at the production patterns that we observe over time. On the other hand, the methodology employed also allow us to rank countries and/or cities with respect to their "fitness" with respect to the "scientific environment" as judged by their ability to produce new knowledge in more complex fields of inquiry.

12. [Ongoing] **Brain Drain, Brain Gain.** We investigate how scientists move and how this affects the scientific gain/drain of countries/institutions. From our analysis, we observe that there exists an overall rapid growth of global scientific activities as well as an increasing trend of international scientific collaborations. We also show how relative rankings among countries - in terms of total scientific production/consumption - change over time and how their ability to retain/attract scientists vary. In addition, we see how scientists exchange affects scientific productivity and whether scientists exchange have different impact across different countries. Lastly, we show how there exists a "preferential flow" between specific pairs of countries.

13. [Manuscript in preparation] **Geography of Scientific Collaboration in Physics.** This study examines spatial-temporal structures of scientific collaboration networks from the American Physical Society publication dataset in the last 50 years. It aims at identifying influence of geographic distance, national boundaries, cultural differences (e.g. language), and frequency of air transportations on the formation of scientific collaborations. The results show that early 1990s are a watershed moment. In the recent 20 years, the frequency of short distance collaborations has decreased and the long distances collaborations have grown dramatically. The comparison of such trends with a null model suggests that there are less and less geographical constraints between scientific collaborations (Figure 1). We also found that the positive correlation between the frequency of air transportation and of scientific collaboration between urban areas steadily increases, meanwhile it becomes less likely that two collaborators come from the same country. This suggests that modern transportation facilities help breaking the national or even cultural boundaries of scientific collaboration. Our findings highlight the importance of spatial-temporal analysis to further address the dynamics of scientific collaboration networks.

14. [Ongoing] **Knowledge Exploration-Exploitation in Science.** Since its introduction by March (1991), the exploration-exploitation dilemma has been extensively studied in many fields and different contexts. Here, we focus on scientists and researchers: in the pursuit of science, how do they behave to expand their knowledge? If one were to follow them on the map of their personal "knowledge space", would one find that they wander erratically, exploring this space or that, once they find a gold vein, they keep digging there? How do they allocate their limited time and resources between these two behaviors? Is there even an optimal allocation? Does this allocation change over their career, perhaps allowing to distinguish the behavior of junior and senior researchers? Have there been changes over time, perhaps pushed by changes in academia and scientific production in the past decades or by a more specialized and competitive world? In this project, we follow individual scientists over all their careers and use their co-authorship (and citation) network to map "what they know", i.e. as a proxy of their knowledge space and of

its evolution over time. Preliminary results show that there is an optimal exploration-exploitation ratio that is chosen by scientists who excel in terms of performance and productivity.

15. [Ongoing] **Research Strategies at the level of Institutions.** This project also falls within that strand of literature interested in the so-called exploration-versus-exploitation dilemma. However, this time our focus is not on the choices made by individual scientists but we look at the behavior and strategies followed by academic institutions. Intuitively, there exists a "tension" between exploration (trying to do research in new fields) and exploitation (keep working in fields in which one is already expert on): on the one hand, exploring is risky and potentially highly rewarding, while on the other hand, exploiting is seen as less risky and capable of yielding lower but safer and consistent benefits. Here we first focus on how to measure and quantify the "amount of exploration" and then we try to assess its impact on performance. We do this at an "aggregate" level, meaning that here the production units are entire (departments of) universities. This analysis sheds light on the (non-monotonic) relationship existing between the exploration-exploitation dilemma (a.k.a. diversification-specialization) and its impact on performance, at an aggregate level. At the same time, it keeps the door open to models and explanations for how to join individual-level behaviors and aggregate outcomes.

16. [Ongoing] **New Data Sources: Microsoft Academic Graph (MAG)**. Along with the Web of Science and APS dataset for research, we have started testing MAG as an additional and complementary source for bibliographic data.

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB

Additional Information

Archival Publications (published) during reporting period:

New discoveries, inventions, or patent disclosures:

Do you have any discoveries, inventions, or patent disclosures to report for this period?

Yes

- No

Changes in research objectives (if any):

None

Change in AFOSR Program Manager, if any:

No change

Extensions granted or milestones slipped, if any:

Not applicable