PRE-EXAMINATION STATEMENT

Mr Pietro Della Briotta Parolo
Analysis of Cumulative and Temporal Patterns in Science

Pre-examiner: Prof. Anxo Sánchez
           Professor of Applied Mathematics
           GISC/Matemáticas, Universidad Carlos III de Madrid

Mr Della Briotta Parolo's Doctoral Dissertation is an article dissertation. It consists of four publications preceded by an in. Two of the publications have already appeared in well-recognized journals (*Phys. Today, J. Informetrics*), which is a strong indication of the quality of the work. The other two articles (Publications III and IV) have not been published yet. However, after careful consideration my assessment is that they are of the same quality as the other two or even more (particularly Publication IV). Therefore, my impression of the overall quality of the work is very good.

Considered as a whole, the dissertation is scientifically correct, starting with clear objectives, namely to contribute meaningful tools in the field of Scientometrics from different viewpoints, both authors and papers. This is true of both the published and the unpublished parts. In this regard, it is my personal opinion that the dissertation contains novel and interesting results that can be useful for the community of researchers in the field. On the other hand, the contribution of the doctoral candidate as described in the dissertation is more than sufficient in terms of research input; in fact, his coauthors are less numerous than for many other researchers in the field.

The main results of the dissertation can be summarized as follows:

1. The time to recognize great scientific discoveries with a Nobel prize is increasing rapidly, either because there is too little scientific progress or too much. The evidence is indisputable and points to a systemic problem for the Nobel Foundation.
2. The time for a paper to reach its maximum number of citations is decreasing in natural years, but roughly constant when time is measured as the number of papers published. The fact that this can be related to an ultradiffusive process indicates that researchers are having problems following the literature because of the increase in publications.
3. Further analysis of the citation patterns in terms of ego-networks for papers confirms the existence of a peak and decay temporal evolution for citations, but provides complementary information regarding how papers get connected to more fields as time passes.
4. Finally, it is shown that a persistent influence index introduced in the dissertation is able to identify Nobel Prize papers as well as other papers

that gave rise to interseting research avenues. At the same time, it allows to quantify how information diffuses between fields.

This is certainly a significant contribution to Scientometrics and it points to a well, logically structured work searching to obtain meaningful results. The candidate has managed to obtain these results by a long process of data analysis which has most certainly not been easy, due to the sheer number of records handled, well into the tenths of millions.

As for the introduction, it clearly places the work in context. I am pleasantly surprised by the vast knowledge of the field demonstrated by the candidate's list of references, almost 200 papers well referenced in the text. In this manner, the candidate is showing the significance of his results in the midst of many other contributions to the field. I have to say here that reading the introduction has made me aware of references I did not know and that can be useful for my work. The depth of the review part is rare for researchers at this stage and is worth praising here.

In spite of my general excellent impression, I would like to recommend a few minor ammendments in the introduction, summarized in the enclosed list by order of appearance in the text. I want to explicitly indicate here that the points I mention below are not important enough to disqualify the work or the presentation in the compendium as a whole, and are to be understood only as recommendations for improvement. In particular, I do not expect a response and a re-review by me is not needed at all.

In summary, this dissertation is certainly sufficient to grant Mr Della Briotta Parolo a Ph D degree. For comparison, in my University, which is among the highest ranked in Spain, and my Department, dissertations with three or four papers published in relevant refereed journals are generally regarded as very good. The quality of the introduction is also comparable to the standards I am used to from Spain, but also from my participation as examiner for Imperial College, University of Oxford, Université de Lausanne, Université de Fribourg, and even for your own institution. The whole dissertation thus fulfills the requirements I would expect and I am happy to recommend that it be granted permission for publication. I congratulate the candidate and his supervisors and advisors for their work. If I can provide any further information or comments, please do not hesitate to contact me.

Signed: Prof. Anxo Sánchez

List of proposed ammendments and revisions

Mr Pietro Della Briotta Parolo
Analysis of Cumulative and Temporal Patterns in Science

- As a general remark, the language needs some revising. The level of English (as far as a foreign speaker like me can say) is good, but there are many long sentences that make it somewhat difficult to follow the discourse at several places. In most cases, this is not a problem, but there are few where it can be and will appear below. In addition, there are a number of typos that should be fixed too. Also, acronyms are not always defined and even two acronyms (PAM and PA) are used for the same thing (I believe). I will e-mail an annotated pdf file with the typos to the candidate to help him locate them.

- On page 16, first line, it is stated that "A simple citation count is a superficial yet quantitative evaluation of the success of a paper and is sufficient to be able to compare [...] publications and scientists." While it is clear from the dissertation that the candidate does not endorse this claim, I believe the sentence should be amended to indicate so, for instance by changing "is sufficient" by is "is deemed sufficient by some" or something to that effect.

- Same page, lines 3-5 on Sec. 2.1, there is a sentence on why the average number of citations should be low that I haven't been able to make much sense of, in fact, the argument in terms of the finite number of references available seems to support the opposite. Please clarify.

- Same page, line 6 from bottom, it is stated that 80% of the papers do not receive more than 10 citations. I wonder how time (which is one of the foci of interest of the dissertation) enters in this assertion. Failed to gather 10 or more citations, after how long? This should be included as it is not the same 10 citations in a year or in 10 years.

- Page 17, third paragraph, lines 4-6, it is simultaneously said that there are universal properties in the citation distributions across disciplines and that the citation distributions are extremely different across fields. This cannot be, it has to be one or the other. Please clarify.

- Page 22, six lines above the figure, as far as I understand the hypothesis posited there is that the focus is constantly shifting towards recent papers, but this seems to contradict the result by Verstak et al. discussed in the preceding paragraphs. A discussion of the connection of the hypothesis with that previous work would be useful here.

GRUPO INTERDISCIPLINAR DE SISTEMAS COMPLEJOS
DEPARTAMENTO DE MATEMÁTICAS
UNIVERSIDAD CARLOS III DE MADRID
Avda. Universidad, 30 – 28911 Leganés, Madrid
Tfno. 916 249 411 Fax 916 249 129 anxo@math.uc3m.es
http://www.anxosanchez.eu

uc3m
GISC

- Same page, in the figure caption the half-life of papers is plotted, but unless I am missing something half-life has not been defined. Please define in the text. Another remark in connection with this figure is that I am not really sure what is the point it intends to make in connection with the text.

- Page 23, first paragraph, there is a discussion of the impact of multi-authorship in a paper on its citations, and it is suggested that having more authors provides a geographical advantage as papers can be promoted in different places. Is it known whether multi-authored papers where all authors are from the same country behave differently from the case in which they are from different countries? That would be grounds for suggesting the geographical mechanism; otherwise, it is probably not a correct explanation.

- Page 26, line 9 from bottom, the sentence ending in "constant r" is unclear. In fact, I am not even sure what "r" means there. Please clarify.

- Page 29, line 8 from bottom, the sentence "first earliest structured books" does not make much sense to me. Please check.

- Page 31, line 9 from top, it is said that sparse matrices are computationally problematic. I understand that the candidate referes that they have many null entries with no information, but in fact that is not a problem, it is an advantage as the candidate describes subsequently by replacing adjacency matrices by adjacency lists. Please rephrase.

- Same page, five lines below, there is a reference to more general networks that is really to multilayer networks. I do not think that multilayer networks are "more general networks", they should be mentioned by name to make clear what this is about.

- Page 32, four lines four bottoms, "citations" is used in the discussion of networks in general, I believe "links" was meant here.

- Page 33, 12 lines from bottom, there is a definition of distance in termes of the shortest path which I cannot understand. The shortest path is already the distance. Are you referring here to the mean over the whole network? In that case please state it explicitly. There it is also said "fully connected". This refers to the complete network. To indicate that the network is not disconnected, "connected" is the word.

- Page 43, there is a mention to ISI as the handler of the Journal Citation Reports (there might be earlier mentions but here is when I noticed). I understand that JCR is now the property of Clarivate Analytics, perhaps this should be the name mentioned here for accuracy.

GRUPO INTERDISCIPLINAR DE SISTEMAS COMPLEJOS
DEPARTAMENTO DE MATEMÁTICAS
UNIVERSIDAD CARLOS III DE MADRID
Avda. Universidad, 30 – 28911 Leganés, Madrid
Tfno. 916 249 411 Fax 916 249 129 anxo@math.uc3m.es
http://www.anxosanchez.eu

uc3m
GISC

- Page 43, paragraph two from bottom, where the SEIR model is described, it is a particularly complex sentence and I believe there may be a verb missing somewhere.

- Page 44, some 15 lines from top, it says "innovative insuccess", do you mean "innovation failure"? Please clarify. Also, the words written a few lines above within mathematical notation should be taken out so they are typeset as regular words.

- Page 48, PageRank is discussed for the second time, I believe you can summarize the discussion and restrict to any aspect that was not mentioned the first time.

- Page 50, line 2 from top, I believe it is not newer papers that inherit, they pass their influence backwards towards earlier papers, if I understood correctly the index. This is actually what I can see from Eq. (4.2). Please clarify.

- Page 52, the paragraph about the results in Publication I does not highlight clearly what its conclusions are, please make it more explicit.

- Page 53, line 4 from top, I understand that you are proposing the ultradiffusive mechanism as responsible for the citation behavior. If that is the case, you should indicate you are proposing it as it is your result.

- Finally, there are several incomplete references. I have spotted a few one and are marked in the pdf file, but there might be more. Please check.