Alexander M. Petersen

Assistant Professor
Department of Management of Complex Systems
University of California, Merced
apetersen3@ucmerced.edu

July 31, 2017

Pre-examination statement on the dissertation manuscript “Analysis of cumulative and temporal patterns in science” by doctoral candidate Mr. Pietro della Briotta Parolo.

This dissertation starts with a well-documented review of the history and recent direction of bibliometric research efforts aimed at understanding the impact of *dynamical* aspects of science and their non-trivial implications. Within this context, the dissertation manuscript summarizes a well-motivated and well-organized series of 4 quantitative analyses.

Each of the analyses address important understudied problems, thereby providing a significant contribution to the literature. As such, it is evident that the doctoral candidate has strong potential to continue along this research stream, demonstrating the skills in research, analysis, modeling, and writing that will help yield opportunities in a wide variety of directions.

Most important results:

Mr. della Briotta Parolo demonstrates in this dissertation the broad skill set required of a doctoral award. First, a knowledge of prior literature (which ironically serves as a topic of his research). Second, the ability to explore and analyze large multi scale datasets, namely the ISI bibliometrics database that was primarily used, in which the candidate accounted for author-level, journal/field-level, temporal-level, and network-level features of the data. Handling this dataset efficiently is not just a matter of computing power, but requires careful algorithmic design and execution. Third, the candidate demonstrates herein demonstrates the technical ability to identify and measure statistical regularities, and to effectively communicate the results both verbally and visually. And finally, the four publications demonstrate a record of persistence, leadership, and scientific contribution.

A brief summary of the results of each publication that are most notable are:

Publication I: A strong demonstration that the Nobel Prize policy that no awards can be made posthumously may need to be reconsidered.

Publication II: How do papers age and how does this aging depend on the definition of time used? This study quantitatively measures and ascribes the functional form (exponential) of the decline in the annual citation rate. Moreover, the study compares two definitions for the units of time, pointing to a severely under-appreciated measurement problem in scientometric studies - how to control for the secular growth of the citation network?

Publication III: This study applies the ego network framework to individual publications. This methodology is well-suited for revealing hidden patterns that underly and help explain the success of highly-cited papers. Per the theme of the dissertation, the study reveals a new time scale in the citation life-cycle of a paper based on dynamic aspects of the citation network. The framework and results provide new avenues to understand long-term scientific impact, as well as to illuminate more nuanced social features, such as the role of self-citation.

Publication IV: This study analyzes the cumulative downstream impact of individual papers, and develops a new measure termed the “persistent influence”. The authors use Nobel papers as a ground truth to provide new understanding of how knowledge spreads via the citation network, and what this may say about emergent trends and the role of inter/multi-disciplinarity in science.

Dissertation Merits:

- Together, the four publications provide a formidable advancement into our knowledge of the role of time in science, refining our understanding of the attribution of credit and the spreading of knowledge. As an expert in this field, I can attest with high confidence to the scientific correctness of the dissertation, which is largely evident in the high level of technical detail in which the studies have been performed.

- An excellent review of the literature as a motivation for the four analyses, each framed well with historical reflection and contemporary context.

- High-quality visualizations integrating eye-catching color schemes, concise multi-panel figures, and network layout techniques.

- Introduction of several new methods: a new definition for measuring time in citation networks (Publication II), the “persistent influence” metric (Publication IV), and a new ego framework for evaluating the correlated citation structure of individual publications (Publication III).

- The dissertation work also addressed topics that have broad appeal, in addition to importance for science policy (Publication I) and the evaluation of research output (Publications II-IV).

- All publications draw on multi-disciplinary literature, are well-written for broad audiences, and demonstrate expertise in a wide range of statistical and computational methods.

- It is clear from the author contributions that the candidate is proficient in analysis and writing, and that the quality of the publications are a valid representation of his research skills.

Dissertation Shortcomings:
The dissertation does an excellent job at reviewing the relevant literature. The principal shortcoming, however, is mainly structural in that the literature review appears to be more prominent than the summary of the dissertation work.

To be specific, in reading in linear fashion, it was not evident until section 5 “Results” that a more detailed discussion of the candidate’s publications I - IV would take place. While reading up to section 5, I was particularly vexed by the overwhelming focus on literature review relative to the candidate’s thesis contributions.

Unless this is a particular formatting policy of Aalto University that I am not aware of, then the dissertation structure should be made even more explicit in the last paragraph of the Introduction section, in which the candidate summarizes Chapters “1-3” (typo > “2-4”; It should also be noted that the chapter numbers indicated in this last paragraph of the introduction do not coincide with the Contents enumeration), but fails to mention the final Chapter 5 - the most important part aside from the appended publications. Thus, it should be stated as explicitly as possible that the purpose of Chapters 2-4 are to provide context for a summary of the main findings of the thesis work in Ch. 5, also specifying that these results are drawn from 4 separate studies.

Indeed, the author may consider finishing this introduction with a summary of the Discussion, e.g. pointing out the fundamental motivation driving the dissertation work, as written concisely in the second to last paragraph of the dissertation, “The research presented in this Thesis presents a diametrically opposed point of view to the matter; science does not represent a static platform for the output of new information, but is rather an ever changing system with sociological, economical and geographical characteristics,…”.

A second shortcoming, again structural, is the abrupt transition to the Chapter 5 “Results”. A brief recap that reframes Publications 1-4 in the context of the literature review of Chapters 2-4 should be considered. Otherwise, this section is more or less a bullet-pointed summary of each paper, and so it is hard to connect the dots in a way. As such, this culminating section does not effectively leverage the grand effort made during the literature review. Thus, an introduction is necessary to firmly establish the main interconnecting themes of time: e.g. life-cycles (of people, ideas, citation histories), burstiness, discovery processes, and even the dimensions/definition of time itself. One final point, which may be a stylistic one, but the Results section lacks citations, e.g. to technical terms such as “ultradiffussive process”.

***

In conclusion, the candidate has presented four finished publications embedded within the framework of a review of the “science of science”. These studies collectively address the non-trivial impact that time has on the measurement and evaluation of scientific careers and publications. The thesis and publications are well-written, and the publications are technically sound and of high (publishable) quality. Together, this body of work demonstrates the candidate’s appropriate level of expertise that is expected of doctoral work.

Hence, I recommend that the thesis be accepted with the opportunity to address the list of minor typographical comments listed in the addendum below (I will leave it up to the committee to forward them to the candidate). In the case that the candidate elects to implement them, they are sufficiently minor that re-examination is not necessary. As such, I recommend that permission be given to publish the dissertation.

Sincerely,

Alex Petersen

Addendum — Minor Editorial Comments:

- Abstract: “The goal of science has always been the one to investigate the world and its phenomena” > “The goal of science has always been to investigate the world and its phenomena”

- page 12: “it is in the nature of science itself to rely one’s work on the top of previous ones and therefore adding” > “it is in the nature of science to build one’s work on the top of previous works, therefore adding”

- page 17: “expected average value” > “expected value” or “average value”; also, if \sigma^2 = \mu then the average value of c_f is exp[3\mu/2]; \sigma^2 = -2\mu will correctly reproduce the statement that the mean value = 1.

- page 20, first sentence: the parameter T is not defined. Also, an additional explanation for the findings of Publication I is that the average human lifetime is increasing. Also, there is a typo “scinetific knowledge”

- page 22: more effort should be made to explain what are the quantities shown in Figure 2.4. As it stands, the reader must infer what the y-axes labels mean.

- page 23: “The same author” > better to indicate which author precisely, as the most recent reference is to “the authors”. Also, typo: “physcosociological”.

- page 24: “the pursue for” > “the pursuit for”.

- page 25: In general, the second paragraph needs a bit more effort defining the quantities introduced. For example, the reader is left wondering what is meant by “citation ages Q(t)”, which could mean multiple things. It should be pointed out that C(.) and T(.) are generic functions, and that the assumption of Hajra was that the attachment rate is separable, which may be a modeling assumption to make the problem tractable, or maybe they are magically separable empirically. Also, a typo: hyphotetized > hypothesized;

- page 26: At the end of the first paragraph, should read c_{t-1}, correct? Also, I did not follow the sentence including “plus a variable *r*, it is normal to expect more recent papers to gather constant r.”

- page 27: typo “to introduce to” > “to introduce two”.

- page 28: “Furthermore, the author” > “Furthermore, the authors”

-page 35: “sharp initial in” > “sharp initial growth in“; and “shown in Fig.3.1”> “shown in Fig.3.3”. Also, it should be pointed out in the legend of Fig. 3.3 that f_i is the analog of modularity within the framework of the ego network. Also, a paragraph should be spent describing what purpose the EN serves, e.g. that it is the realistic -local- perspective of a given node representing the information that it might use in basic decision processes.

-page 36: Missing a period at the end of the first paragraph. Figure 3.3 label: “approx” > “\approx”.

- page 37: The sentence “Sociological considerations [100] can support the hypothesis that preferential attachment method is the the driving force only of collaboration only for scientists in the middle of the career (thus also in the middle of the distribution), with the tails of the distribution being dominated by either established scientist, who don’t require to build up their network anymore, or newcomers who instead fail to act as attractors in the network.“ is too long, and confuses terms (e.g. preferential attachment) previously defined with respect to the citation network, now with respect to the coauthor network. A clarification that these concepts translate to the context of co-authorship networks would help. Also, there is a typo in Reference 100 and 110.

- page 40: “paper based” > “paper-based”; also, it is worth mentioning that self-citations likely play a strong role here, certainly in the case of the example EN shown in 3.4. Does the pattern change dramatically if self-citations are ignored?

- Page 45: There should be a paragraph break before “In Publication IV” and more time should be spent discussing the results and their implications. For example, the statement “A renormalization of time similar to the one in Publication I shows that the trend of increased interdisciplinarity is actually reversed” should be more thoroughly explained given the striking nature of the statement/result.

- Page 46: “decresing” > “decreasing”. Also, the sentence “We can see that the coloring order between the two columns is reversed, indicating that also for subfields and journals are on average the same pattern as for the fields applies.” is confusing, and in general, panels c-f aren’t very informative with respect to what is discussed on page 45.

- Page 49: “matrix of the WWW, which proves that the PR is a centrality measure.” This sentence is a bit strange, suggesting that A is defined only for the WWW and that the previous argument is a mathematical proof. Also, at the bottom of the page “which he knowledge” > “which the knowledge”.

- Page 50: The axes labels for Fig. 4.1 could be improved to better match the typeface used in the text.