Fraud Detection in Statistics Education based on the Compendium Platform and Reproducible Computing

Patrick Wessa
K.U.Leuven Association
Lessius Dept. of Business Studies
Belgium

Bart Baesens
K.U.Leuven Association
Faculty of Business and Economics
Belgium

Abstract

This paper focuses on a newly developed method to detect fraud in empirical papers that are submitted by students. The proposed solution is based on the Compendium Platform and Reproducible Computing ([21], [18], [17], [20]) which allows the educator to build e-learning environments that are embedded in the pedagogical framework of social constructivism ([16], [15], [12]) and which can be shown to be effective in terms of non-rote learning of statistical concepts [19].

The paper addresses the technological aspects of the proposed fraud detection system, ways to discriminate between various types of fraud (plagiarism, free riding, data tampering, peer-review cheating), and the pedagogical issues that result from its implementation (responsibility, non-rote learning). Finally, the first experiences about the implementation of the proposed technology in three undergraduate statistics courses (with a large student populations) are discussed, and used to recommend paths for future research & development.

Acknowledgments

This research was funded by the OOF2007/13 project of the K.U.Leuven Association. I would like to thank Ed van Stee for his excellent work in programming substantial parts of the Compendium Platform.

1. Introduction

In a recent editorial of the journal *Research Policy* the problem of plagiarism and the inability of peer review to detect plagiarism was clearly illustrated and summarized as follows [4]: *The fact that academic misconduct on this scale has gone unchecked over such a prolonged period raises serious issues about the efficacy of the processes used to police the conduct of researchers. ... a measured degree of vigilance and a greater willingness to pursue any well-founded suspicions of research misconduct are required by editors, referees, publishers and the wider academic community if the scourge of plagiarism is to be kept at bay.*

If this is true for the research community then the problem of fraud detection in education is not only relevant but also very challenging. In particular, we believe that it is difficult to detect fraudulent activities that are related to statistical analysis because of a variety of reasons, such as the following:

- the data under investigation might not be readily available
- the software that is needed to verify the analysis might not be available
- the computation might not be reproducible because it is obfuscated (for instance when the underlying computational parameters are not explicitly defined)

The difficulties that we encounter to detect statistical fraud is therefore closely related to the problem of irreproducible research as described in [6] and [2]. Many solutions have been proposed ([3],[13], [14], [1], [5], [8], [9]) but were not implemented in statistics education due to a series of practical and technical reasons [20].

With the introduction of our newly developed Reproducible Computing technology (which is hosted within the so-called Compendium Platform and which was built upon the R Framework for statistical computing [17]) these problems have been solved [20], [19]. In addition, it is now possible to accurately measure the actual - rather than self-reported - learning activities that are related to statistical computing [20]. This is not only important to gain a better understanding of learning processes and their relationships with computing and learning technology. It is also a conditio sine qua non for improving fraud detection and prevention as will be illustrated in the following sections.

2. Statistics Education Data

A large amount of data was collected within the context of a series of experimental, undergraduate statistics courses in an academic business school in Belgium. One particular course was selected to create the computations and illustrations in this paper - the other courses exhibit results that are quite similar. The selected course had a total student population of size 103 (after elimination of drop-outs).

The course contained a wide variety of statistical techniques and methods such as: probability, discrete and continuous distributions, descriptive statistics, explorative data analysis, hypothesis testing, multiple linear regression, and univariate time series analysis. A total of 73 different types of statistical techniques were covered in the course with a large variety of model parameters.

For each technique, students had one or several web-based software modules available which are based on the R Framework and are available free of charge at http://www.wessa.net/. The R Framework allows educators and scientists to develop new (and easily maintain) tailor-made statistical software and at the same time the end-user is able to change the underlying source code and improve the software [17].

The main sections of the statistics course were built around a series of research-based workshops that require students to reflect and communicate about a variety of statistical problems, at various levels of difficulty. The workshops have been carefully designed and cannot be solved without additional information that is provided within the Virtual Learning Environment or by the educator.

Students were required to perform detailed peer reviews of about 5-7 submissions from other students. Peer review of statistical papers (by students) is made possible by the fact that our Reproducible Computing technology [20] is easily accessible and does not require the reader to download or install anything. The reader/reviewer is not even required to understand the underlying R code of the statistical analysis to reproduce or reuse the computations - a simple click on a hyperlink is sufficient to view all the meta data that is associated with a statistical computation in a web page. The meta data web page contains a button that allows the reader to recompute the analysis in real-time, based on the online server-based statistical software that was made available. Whenever a reader/reviewer reproduces a computation, there are many ways in which the computation can be reused - for instance, with different parameters, methods, or data. In fact, all features of the R Framework for statistical computing [17] are available within a few mouse clicks. This allows any student to easily, and quickly review the work of peers without any technical problems (such as installation, finding the right parameters, etc...).

Even though students had to assess the submitted workshops and give them a score, the peer review was not only intended to be an evaluation method - it also enabled them to provide feedback, learn from mistakes made by others, communicate solutions about a variety of problems, and provide an incentive in the form of encouragement to fellow students. The assessment/review requirement was an integral part of the actual learning process of students and was intended to nurture their attitudes towards critical thinking and scientific honesty.

As one might have noted, this feedback-oriented process is similar to the peer review procedure of an article that is submitted to a scientific journal. The process of peer review is an important aspect of scientific endeavor, and may help us in achieving learning goals with respect to attitudes (through peer review experiences) and skills (through constructivism [16], [15], [12]). Moreover, in previous research it was found that there is strong empirical evidence that the use of Reproducible Computing is related to non-rote learning of statistical concepts which is measured by objective exam questions [19].

Each student submitted a total of 15 different workshop papers (one or two papers per week) which were subjected to peer review. Every submission was assessed with respect to 3-9 criteria. For every graded criterion students had the ability to provide verbal feedback to the other student. The otherwise time-consuming administration of the Peer Assessment procedure was automatically performed by the use of the Virtual Learning Environment called Moodle [10] which is freely available and which has been designed with a constructivist, pedagogical philosophy in mind. The educator graded the quality of the verbal feedback messages that were submitted to other students which provided them with a strong incentive to do a good job with their peer reviews.

More importantly, students were required to assess the authorship of computations. For instance, if a submission contained irreproducible computations or if the author of the computation was not properly identified, they had to give a very low grade. All forms of fraud (such as plagiarism and free riding) had to be clearly identified and reported by students. Quite a lot of fraudulent cases (in early papers) were detected and students were well-aware of the importance of honesty and reproducibility. As the course semester progressed, fewer cases of fraud were detected/reported which may be due to a learning process. However, in the following section it will be illustrated that - even under these intensive review conditions - fraud was not adequately detected by peers and that additional information (about the actually underlying computations) is required to improve the detection process.

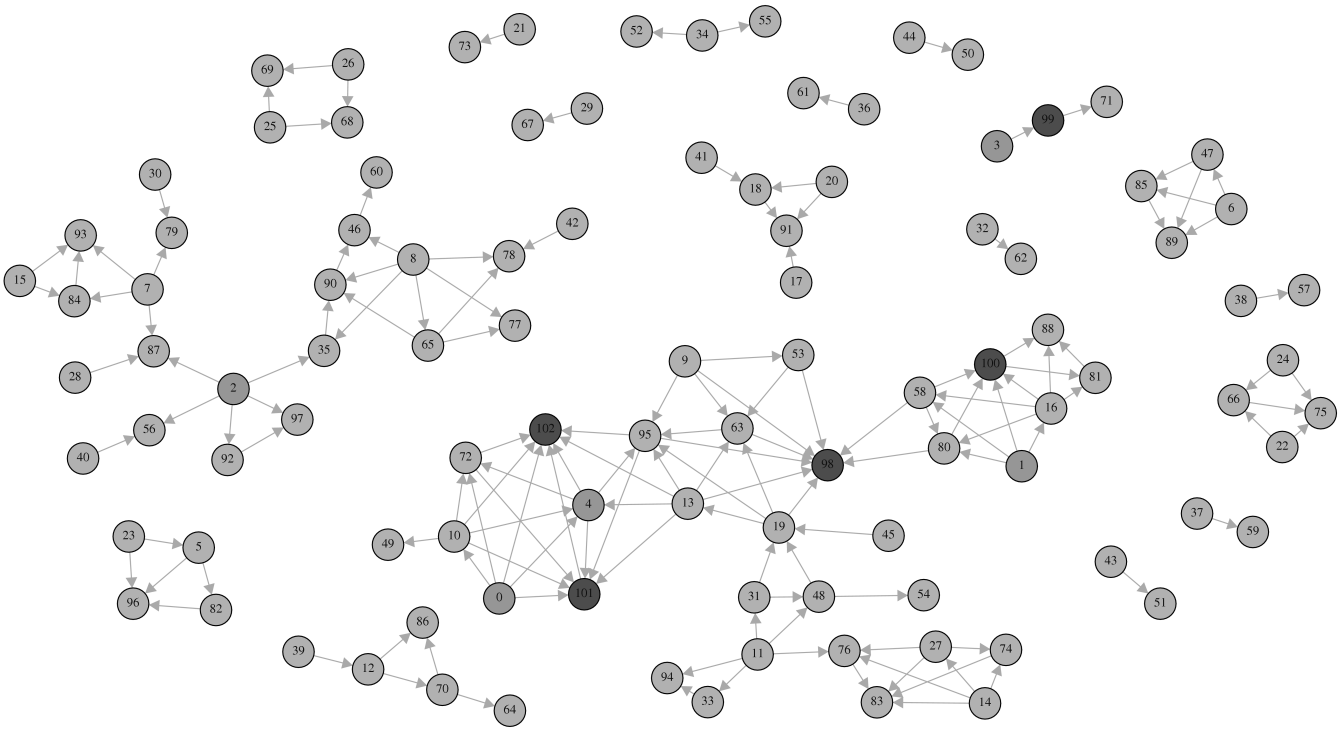

Figure 1. The Sociogram based on Paper Submissions and Reproducible Computing

3. Fraud Detection

3.1. Social Networks

In general, a Social Network is defined as the set of N students S_n (so-called "actors") and their pair-wise relationships which can be described by (so-called $N \times N$ "sociomatrix") $Y \equiv [Y_{n,m}]$ where

$$Y_{n,m} = \begin{cases} 1 & \text{relationship from actor } n \text{ to } m \\ 0 & \text{otherwise} \end{cases}$$

In this case, the sociomatrix is constructed on the basis of a network of archived computations which are used in K assignments. Every student $m = 1, 2, ..., N$ works on each assignment $k = 1, 2, ..., K$ during a pre-specified assignment period. Each submitted paper $P_{k,m}$ contains a number ($N_{k,m}$) of links that refer to archived computations $C_{k,i,m}$ for $i = 1, 2, ..., N_{k,m}$ which was created and archived by student $S_n \equiv \sigma\left(C_{k,i,m}\right)$.

In a traditional fraud detection system, one would compare the phrases of papers $P_{k,m}$ and $P_{k,n}$ for $n \neq m$ in order to detect plagiarism. The proposed system is fundamentally different because it is not based on the wording but on the social relationships between students. A social relationship $S_n \rightarrow S_m$ is created whenever student m borrows a computation from student n.

Social networks often display a (very) large size [11] and have a complex structure which requires computational considerations and relationship algebra [7]. Within the context of peer influence groups in particular, it has been shown that such large networks can be simplified through special methods that identify clusters and reduce complexity [11]. Instead of using such sophisticated techniques our approach is to simplify the sociomatrix - with the purpose to obtain an aggregated overview - by defining the net number of lent computations which is computed as follows: $L_{n,m} = \sum_{k=1}^{K} \sum_{i=1}^{N_{k,m}} \iota\left(\sigma\left(C_{k,i,m}\right) = n\right)$ for $n, m = 1, 2, ..., N$ where $\iota\left(x = n\right) = \begin{cases} 1 & x = n \\ 0 & x \neq n \end{cases}$.

The underlying assumptions of this simplification are as follows:

- fraudulent activities are related to free riding
- free riders are net borrowers of computations
- all computations are equally important
- the number of self-created computations (whether cited or not) is unrelated to fraud

The so-called "edges" of the sociogram can now be easily obtained:

$$Y_{n,m} = \begin{cases} 1 & L_{n,m} - L_{m,n} > 0 \\ 0 & \text{otherwise} \end{cases}$$

which yields - in our experience - a convenient number of relationships in the social network even when the student population is large.

Without loss of generality it is possible to sort students S_n according to their total lending score $T_n \equiv \sum_m^N L_{n,m}$ (in descending order) such that any relationship $S_n \rightarrow S_m$ can be interpreted in terms of $|n-m|$. The larger the difference $|n-m|$ becomes, the more likely it is that student S_m is a free rider. Also, the number of students from whom student S_m borrows, might be related to the extent or severity of fraudulent activities. Obviously, it is impossible to set a universal standard or benchmark by which both measures can reliably indicate the presense of fraud. A statistical model - which is beyond the scope of this paper - would be required to estimate the "optimal" threshold value that is to be compared against $|n-m|$ in order to predict fraud.

For the sake of illustration, the above model is applied to the data that was described in the previous section. The resulting sociogram can be seen in Figure 1 in which one can observe how the students are socially related to each other. It is clearly seen that students are clustered and that each individual student falls either into a small group (of about 2-5 students) or into one (of two) large networks of "collaborating" students.

3.2. Free Riding

Based on a careful analysis of their papers, students S_{102} and S_{101} have been identified as free riders. They both borrowed a substantial proportion of computations from other students and cited them as their own - much to our surprise, this was also the case for the final paper. If one examines the textual content of the papers there is nothing which leads the reader to suspect them. Peer review never indicated any problems and their aggregate peer assessment scores were among the highest of all students.

From Figure 1 it is clear that both students are in close relationship with students S_0 and S_4 which both have high peer assessment scores. These observations reveal three important issues:

- the fraudulent students managed to obtain some form of social connection with S_0 and S_4 which they obviously abused
- our model correctly predicts the fraud because the relationships between the four students exhibit very high differences in index numbers $|n-m|$
- the fact that S_{102} and S_{101} had excellent assessment scores is mainly a caused by the high scores that were righteously attributed to S_0 and S_4 and the fact that the fraud remained undetected during the peer review process

The first and third observation may have psychological and pedagogical consequences. The second observation indicates empirical support for the proposed approach.

Other students engaging in fraudulent activities can also be easily identified: S_m for $m = 97, 98, 99, 100$ are all free riders. Among those, only student S_{98} did not have a high aggregate peer assessment score which may be explained by the fact that S_{98} does not have a direct relationship with a student who achieved a top assessment score (or had very low index n). More importantly, none of the mentioned students was identified as a fraud by means of peer review which gives us reason to believe that social mapping of student interactions (based on Reproducible Computing) may become a necessity in future fraud detection systems that are used in connection with empirical research and statistics education.

3.3. Obfuscation and Data Tampering

Some students understood the consequences of Reproducible Computing technology in terms of fraud detection and tried to obfuscate their analysis by manually reproducing computations of peers and submitting the results to the archive anonymously. One example is student S_{100} who referenced anonymously submitted computations in his paper. This type of fraud can be detected by comparing the archived texts and pictures of the anonymous computations with the entire repository of computations. For this purpose several existing technologies can be used, among which: a search engine and an algorithm that produces an indexable fingerprint (of any picture or text output). Of course the educator might also treat any anonymous computation as untrustworthy and penalize the student accordingly.

Some other students simply reproduced the "borrowed" computations from other students by using the automatic reproduction feature of the Compendium Platform and archived the result (under their own account or anonymously). This attempt to obfuscate the original author of the analysis is however futile. The reason is that for every archived computation the Compendium Platform also stores the parent-child relationships between computations. This means that a relationship is created whenever a computation A is reproduced or reused as a new computation B: the system automatically identifies computation A as the parent of B. It only requires the educator to have the sociogram recomputed, redefining the ownership function $S_n \equiv \sigma\left(C_{k,i,m}\right)$ such that n becomes the index of the student who owns the original/initial parent computation.

Similarly, any attempt to tamper with the data in order to "make a model work" is easy to detect. This is because the Compendium Platform will always compare the parent and child computations in order to identify any change in parameters, in the data, or the underlying R code. Moreover the educator is able to archive a computation that includes the data set of interest. Students are required to reuse the original computation in their assignments, hereby ensuring

that the data cannot be tampered with in any way.

4. Conclusions

Fraud detection is an important and difficult challenge. Reproducible Computing is not only promising in helping students to learn statistical concepts, it also allows us to measure learning activities in ways that was never available before. Social interaction and collaboration can be measured and studied. Some types of social networks may be beneficial (in terms of learning outcomes) while other types allow us to identify frauds that seem to be undetectable within the traditional paradigm of peer review. The bottom line is that we cannot rely on the output alone - the actual research-related learning activities (computations) must be measured and monitored. To paraphrase the Editorial of the journal *Research Policy*: *...a measured degree of vigilance and better technological tools to pursue any well-founded suspicions of research misconduct are required ... if the scourge of plagiarism is to be kept at bay*.

References

[1] J. Buckheit and D. L. Donoho. *Wavelets and Statistics*, chapter Wavelab and reproducible research. Springer-Verlag, 1995.

[2] J. de Leeuw. Reproducible research: the bottom line. In *Department of Statistics Papers, 2001031101*. Department of Statistics, UCLA., 2001.

[3] D. L. Donoho and X. Huo. Beamlab and reproducible research. *International Journal of Wavelets, Multiresolution and Information Processing*, 2004.

[4] Editorial. Keeping plagiarism at bay - a salutary tale. *Research Policy*, 36:905–911, 2007.

[5] R. Gentleman. Applying reproducible research in scientific discovery. BioSilico, 2005.

[6] P. J. Green. Diversities of gifts, but the same spirit. *The Statistician*, pages 423–438, 2003.

[7] J. I. Khan and S. S. Shaikh. Computing in social networks with relationship algebra. *Journal of Network and Computer Applications*, 31:862–878, 2008.

[8] R. Koenker and A. Zeileis. Reproducible econometric research (a critical review of the state of the art). In *Research Report Series*, number 60. Department of Statistics and Mathematics Wirtschaftsuniversität Wien, 2007.

[9] F. Leisch. Sweave and beyond: Computations on text documents. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, 2003.

[10] Moodle. A free, open source course management system for online learning. In *http://www.moodle.org*, 2008.

[11] J. Moody. Peer influence groups: identifying dense clusters in large networks. *Social Networks*, 23:261–283, 2001.

[12] L. Moreno, C. Gonzalez, I. Castilla, E. Gonzalez, and J. Sigut. Applying a constructivist and collaborative methodological approach in engineering education. *Computers & Education*, 49:891–915, 2007.

[13] R. D. Peng, F. Dominici, and S. L. Zeger. Reproducible epidemiologic research. *American Journal of Epidemiology*, 2006.

[14] M. Schwab, N. Karrenbach, and J. Claerbout. Making scientific computations reproducible. *Computing in Science & Engineering*, 2(6):61–67, 2000.

[15] E. Smith. Social constructivism, individual constructivism and the role of computers in mathematics education. *Journal of mathematical behavior*, 17(4), 1999.

[16] E. Von Glasersfeld. Learning as a constructive activity. In *Problems of Representation in the Teaching and Learning of Mathematics*, pages 3–17. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[17] P. Wessa. A framework for statistical software development, maintenance, and publishing within an open-access business model. *Computational Statistics*, 2008.

[18] P. Wessa. *Free Statistics Software (online software at http://www.wessa.net)*. Office for Research Development and Education, 1.1.23-r2 edition, 2008.

[19] P. Wessa. How reproducible research leads to non-rote learning within a socially constructivist e-learning environment. In *Proceedings of the 7th European Conference on e-Learning*, Cyprus, 2008.

[20] P. Wessa. Learning statistics based on the compendium and reproducible computing. In *Proceedings of the World Congress on Engineering and Computer Science (International Conference on Education and Information Technology)*, Berkeley, San Francisco, USA, 2008. UC Berkeley, San Francisco, USA.

[21] P. Wessa and E. van Stee. *Statistical Computations Archive (online software at http://www.freestatistics.org)*. K.U.Leuven Association, Belgium, 2008.