Exploring Social Networks in Reproducible Computing and Collaborative Assignments

Patrick Wessa
K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium
patrick@wessa.net

Abstract: Social constructivism and computer-assisted learning has received a great deal of attention in the pedagogical and technological research literature. However, there is almost no knowledge about the effect of social networks (c.q. interaction between students) in computational assignments and the educator's role as a provider of exemplary cases. Most studies have exclusively employed survey data to investigate social aspects of educational computing which can be shown to be highly misleading – if not biased. Some studies have tried to explore social networking based on objective measurements of forum data (c.q. discussion threads in learning environments) – the fundamental problem with these studies however, is the lack of content-related meta information (there is no information about the content of the discussion unless all posts are coded by the researcher).

In contrast to existing literature, this paper presents an illustrated exploration of an educational database which contains objective measurements of meaningful, social behavior in statistical computing based on several statistics courses with large student populations. All communications between students are uniquely identified by their statistical and educational context which implies that crucial types of objective meta information about the meaning of those messages is available.

The main emphasis is on the following educational aspects: collaboration between students, competition between groups of students, and the usefulness of worked-out examples that are provided by the educator. The aim of this paper is to show that the newly developed Reproducible Computing technology provides us with new ways to research social networking/interaction in assignment-based and constructivist learning. In addition, it is shown that a series of new research questions arise from the explorative data analysis of the measurements that were collected.

Keywords: reproducible computing, social networking, collaboration, constructivism, assignments

1. Introduction

The educational research community has shown great interest in the pedagogical paradigm of social constructivism (Eggen and Kauchak 2001, Smith 1999, Von Glasersfeld 1987) and the mapping of computer technologies that support various types of pedagogical approaches (Conole et al. 2004) and constructivism in science education in particular (Moreno et al. 2007).

With the prospect of computer-assisted learning environments there are new opportunities to measure certain types of social interaction. These relationships are often represented in the form of a graph in which the students are related to each other. A typical example is the analysis of forum data with hierarchical threads of communication between students. The problem with such data however, is the fact that the messages in a forum are not accompanied by meta data that accurately describes the nature or purpose of the message that was posted by a student. A forum message may be relevant in some pedagogical sense but there is no way to discriminate between fundamental discussions about the core subjects of the course and plain small-talk unless the messages go through the difficult and time-consuming process of manual coding. In Dennen (2008) this problem is well described: *Discussion is a required component of many Web-based classes, but do we really know its value or contribution to learning? Students may be graded for participation, and number and length of posts may be counted by those evaluating or researching online classes, but all too often the assessment and analysis methods that we use fail to provide us with data that indicate learning took place through participation in online discussion.*

In addition, social networks often display a (very) large size (Moody 2001) and have a complex structure which requires computational considerations and relationship algebra (Khan and Shaikh 2008). This makes the analysis of social interactions even more challenging and leads us to the conclusion that various filters should be employed in order to be able to explore social interactions that are both meaningful (from the pedagogical perspective) and important (exerting a substantial impact on groups of students rather than on one individual) as will be explained in section 2.

Moreover, there seems to be no literature about the role and effect of computational/analytical assignments or learning activities within the context of social collaboration and constructivist learning. This comes as a surprise because typical statistics courses (and courses that are based on empirical research) often involve substantial amounts of statistical computations which is highly time-consuming for the students and the educator. What this means is that in empirical research-related education the most important activities (c.q. statistical computing) are unmeasured and have an unknown relationship with the social aspects of learning. It looks like only survey-based data about computer use can be used in educational research. This is problematic because reported data can be shown to be highly misleading when compared to actual observations (Wessa 2008b). The introduction of the new Reproducible Computing technology however promises to change that by allowing educators and researchers to objectively measure actual learning activities (such as statistical computing and communication about computational results) as is explained in section 3.

A comprehensive dataset was collected in order to provide a first glimpse inside the social networking behaviour that emerges when students engage in learning activities based on statistical computing and analyses. Section 4 describes the data and the learning environment in which the measurements were collected.

Finally, section 5 discusses the social networks (section 2) that can be explored when Reproducible Computing technology (section 3) is implemented in statistics education (section 4). The purpose of this paper is to illustrate new opportunities and to describe a series of new challenges for pedagogical researchers.

2. Social Networks

The Social Network that is used in this paper is based on a mathematical model in which there are N students S_n with social interactions between all possible pairs S_n and S_m. The set of all combinations of pair-wise relations is represented by the sociomatrix $[Y_{n,m}]$ with N rows and N columns. The cells of the sociomatrix contain a binary value $Y_{n,m}$ that indicates if there is a relationship from S_n towards S_m. For the sake of illustration, suppose that we are interested in social interactions based on peer assessment and that there are only $N = 3$ students. In this case the sociomatrix might take the form of Table 1.

Table 1: Example of a sociomatrix

row S_n "assesses" column S_m	S_1	S_2	S_3
S_1	1	1	0
S_2	0	0	1
S_3	0	1	0

In this example S_1 is the only student who performed a self assessment. In addition, S_2 and S_3 assessed each others work (but now their own). Also it can be seen that student S_2 was assessed by S_1 but S_2 did not assess S_1.

The sociomatrix contains interesting information about the social interactions between students within the context of a specific type of learning activity. Various types of social patterns may emerge in the sociomatrix which may help us to gain a better understanding about the the role and importance of (computer-assisted) social interactions in effective learning.

The row sums of the sociomatrix are defined as $R_n = \Sigma_m Y_{n,m}$ for $n = 1, 2, ..., N$ which can be interpreted as the assessment impact of student S_n. The column sums are $C_m = \Sigma_n Y_{n,m}$ for $m = 1, 2, ..., N$ which measures the number of assessments that have been generated about S_m.

To display the sociomatrix in a graphical form the relationships $Y_{n,m}$ are transformed to a list with pair-wise associations: $S_1 \rightarrow S_1$, $S_1 \rightarrow S_2$, $S_2 \rightarrow S_3$, $S_3 \rightarrow S_2$ which can be displayed as vertices that are connected with arrows between them. Such a graph is called a sociogram and serves as an easy tool to discover/explore the social patterns between students. A combination of statistical algorithms and manual interaction (based on an interactive, graphical user interface) is needed to make the sociogram meaningful. The reason for this is that the position of the vertices – while not representing

any real information – has important repercussions for our ability to visually identify patterns of interest. There are an infinite number of ways (combinations of x-y coordinates) that can be used to represent the relationships from Table 1. Two possible illustrations (including a “good” and “bad” one) are shown in Figure 1. The “bad” example is mathematically equivalent to the “good” example – however, most human observers prefer the “good” sociogram because it highlights the fact that student S_1 plays a special role within the context of assessment interactions.

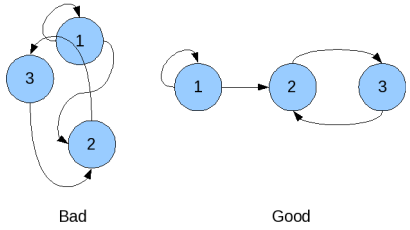

Figure 1: good and bad example

The complexity of social networks grows quickly when the number of students and social activities increases. Therefore it makes sense to introduce simplifications that allow us to focus on important aspects. This can be achieved by eliminating the details that obfuscate the patterns of interest by introducing simplification rules, such as:

- removing all reflexive relationships (we are only interested in social interaction which – by definition – involves more than one individual)
- replacing all measured interactions by “net effects” (eliminate the symmetry in pair-wise relationships)
- displaying the vertices with a meaningful rank order

In order to obtain “net effects” it is possible to use the difference between the number of social actions between a pair of students instead of the absolute number. For instance, if S_2 assesses S_3 for a total of 4 assignments and S_3 assesses S_2 only once then S_2 would become the assessor of S_3 but not vice versa. In this case we would define $Y_{2,3} = 1$ and $Y_{3,2} = 0$. This rule can be applied for each pair of associations between S_n and S_m which greatly reduces the network's complexity (see Figure 2) and improves the readability of the sociogram.

In order to add relevant – and easily readable – information to the network it is possible to assign rank numbers to each of the vertices (students). This can be done without loss of generality – for instance if the “total number of assessments” is used to rank the students in the example (and if the other simplification rules are applied as well) then the results that are displayed in Figure 2 could be obtained.

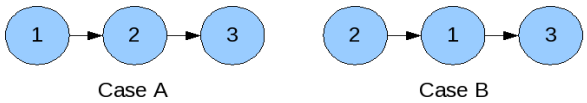

Figure 2: simplified sociogram with additional information based on rank order

In both cases the students are represented in the same location (left, middle, and right position). In case A the student in the left position has generated the largest number of assessments. In case B however the left student has fewer assessments than the one in the middle position. The point of displaying the rank orders (instead of randomly assigned index numbers) is the fact that is possible that the social interaction pattern is somehow related to the rank order. The importance of such relationships (between rank and structure) will be illustrated in section 5 based on empirical evidence.

3. Reproducible Computing

In the academic community, it has been well documented that it is nearly impossible for a reader of an empirical research paper to adequately reproduce – let alone reuse – the research results that are presented (Schwab et al. 2000, de Leeuw 2001, Green 2003). The are plenty of reasons why this is

the case (Peng et al. 2006, Koenker and Zeileis 2007) and only few technological solutions (Donoho and Huo 2004, Gentleman 2005, Leisch 2003) have been proposed – alas, none of these have been shown to be of practical use in education. In order to solve these problems, a new technology was developed which empowers students to easily reproduce and reuse computations that are presented in papers that were created by the educator or other students without the need to download and install anything on the client machine and with no requirement to understand the technicalities that are associated with Reproducible Computing technology or the underlying statistical computations (Wessa 2008a, Wessa 2009a).

While the role and effect of Reproducible Computing for education is clear (Wessa and van Stee 2009b) there is another important advantage which relates to the fact the underlying technology allows us to measure all computer-assisted learning activities, including (but not limited to) the statistical computations that are generated and the associated communications between students (in the context of peer review). The technology does not only keep track of the various actions of individual students but also stores the dependencies thereof. Therefore it is possible to use Reproducible Computing technology for the purpose of educational research and the analysis of social interactions between students based on objectively measured data that are related to computer-assisted learning activities.

More precisely, each computation $C_{h,n}$ (for h = 1, 2, ..., H_n) that is archived by student S_n is available to other students S_m (for $m <> n$). There are (at least) two reasons why a student S_m might have an interest in computations from peers (or the educator):

- to challenge an analysis because it is thought to be faulty or non-optimal
- to experiment with, and learn from the analyses of others

The first reason is related to the process of peer assessment in which student S_m plays the role of (anonymous) reviewer. In this case the relationship $S_n \rightarrow S_m$ means that S_n "is reviewed or challenged by" S_m. The second reason is associated with collaboration and learning from selected computations which are believed to be of interest – for instance, when the analysis which was applied by S_n on a dataset D_n can be reused by S_m and applied to a similar problem with another dataset D_m. In this case the relationship $S_n \rightarrow S_m$ means that S_n "has an impact on" S_m. In this context it is not only interesting to look at the number of arrows that leave from student S_n but also the hierarchical structure of impact which reflects the ways in which good ideas propagate through the social network (within the socially constructivist course).

4. 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. The courses were developed over a time frame of several years (starting in October 2003) but the data were collected from the courses in the last two years (2007 – 2008).

Each course treated a wide variety of statistical techniques and methods including (but not limited to): explorative data analysis, hypothesis testing, multiple regression, and time series analysis. For each technique, students had one or several web-based software modules at their disposal which are freely available (http://www.wessa.net/). In addition, students were able to make use of an online repository of archived computations that can be easily reproduced and reused (http://www.freestatistics.org/).

In 2007 students worked on 9 weekly assignments during the semester and in 2008 there were 13 sequential assignments. Each week there were one or two lectures in which students received information about the past assignments: the educator presented sample submissions, and illustrated good approaches to solve the problems as well as commonly made mistakes. Starting with the second week, students were required to work on the next assignment and (at the same time) participate in review activities based on past submissions from peers. In 2008 students were able to consult the submissions and computations from the previous year, and use them as a guideline. Of course, the assignment problems in the second year were more challenging than in the first because students already had examples at their disposal.

As explained in section 3 the dependencies between computations are stored in the computational repository and can be used to determine "impact" dependencies between students $S_n \rightarrow S_m$. These relationships can not only be measured between students of the same year but also across years. If a

student S_m (in 2008) is reading a 2007 submission from another student S_n then it is possible for S_n to have an "impact" on S_m even if both students have never met each other. This (unusual) definition of social interaction is on purpose because the underlying philosophy of the mentioned statistics courses is to reflect the real world of empirical, academic research – researchers can be influenced by their colleague's work through conference presentations (or face-to-face discussions) but also through articles that are published in journals. In this analogy, the selected paper submissions from 2007 that are made available to the students in 2008, played the role of journal articles.

Both types of impact were measured and are contained in the analysis that is portrayed in the next section. However, in order to obtain a human-readable graph, various filters were applied to the data prior to visualisation:

- only computations that were related to the selected courses were considered – the computations that were contained in documents that were not part of the course were eliminated
- only "net effects" were used (see section 2)
- only the relationships for which the number of reproduced computations was greater than 3 were considered because the number of edges (arrows in the sociogram) would otherwise obfuscate the important patterns – as a consequence, each relationship $S_n \rightarrow S_m$ implies that S_n has "a consistent impact" on S_m (occasional impacts have been removed)

5. Results

Figure 3 shows the sociogram for the dataset from section 4. The sociogram was computed with *R* (R Development Core Team 2008) and the *igraph* package (Csardi and Nepusz 2006). The rank order (c.q. number displayed in the vertices) is based on the total number of times that computations from S_n were reproduced by someone else (= R_n) which can be interpreted as the "total impact" of student *n* (note: high rank orders correspond to high R_n values).

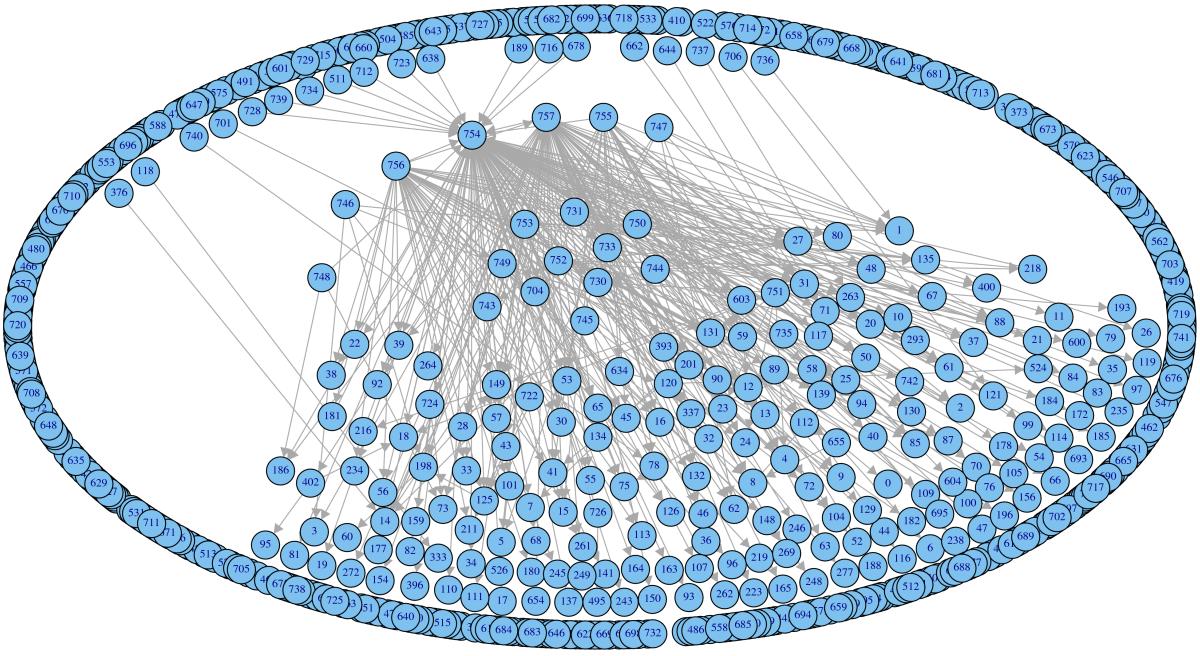

Figure 3: Sociogram based on Reproducible Computing in Statistics Education

The sociogram was obtained with the so-called "Spring Embedder" algorithm as implemented in *igraph* (Csardi and Nepusz 2006) and with the help of manual re-positioning of the vertices, in order to improve the readability of the graph. Figure 3 exhibits some interesting features about the computing-based, social interactions between the students:

- The ellipse contains the students that did only have occasional relationships with other students (in each two-by-two interaction the number of computations is lower than 4). Close observation of the numbers reveals that there are quite a few students with large rank orders who are located on the ellipse. Most of them have generated (a large number of)

computations that were frequently reproduced but did not propagate to new important findings. It is fair to assume that the computations from these students are less important from a social networking point of view.

- In the bottom area (inside the ellipse) there are many students with low rank numbers and only one inbound arrow. These are students who decided to reproduce results (many times) from one particular (influential) source. Most of them are "net importers of ideas" and have (almost) no impact on other students.
- In the middle of the ellipse there is a "central cluster" of students (S_{743}, S_{749}, S_{704}, S_{753}, S_{752}, ...) that have multiple inbound arrows and a large number of outbound arrows. These students seem to play an important role in the dissemination of ideas.
- Between the "net importers" and the "central cluster" there are students with multiple inbound arrows (they use different sources) but have (almost) no impact on other students – most of them have low rank numbers.
- There are 5 students (S_{746}, S_{756}, S_{757}, S_{755}, S_{747}) with multiple outbound arrows and no inbound arrows (except for S_{757} with one inbound arrow). These students seem to be "famous" and have a "high impact" on the entire student population (either directly or indirectly).
- Student S_{754} is exceptional because of the many inbound and very large number of outbound arrows. This student relied heavily on the work that was created in 2007 and had a huge impact on peers in 2008. In this sense this student played the role of a "transmitter" from one period to the next.
- Students in the top area of the ellipse (S_{701}, S_{728}, S_{739}, ..., S_{737}, S_{706}, S_{736}) have a single outbound arrow which influences a student with high impact. Maybe these students play the role of "originator" of important ideas that were picked up by the more popular (high-impact) students.

The above findings are rather robust with respect to the filter that was used to reduce the number of arrows. In other words, if the sociogram is computed with a threshold of 2, 4, or 5 computations per arrow instead of 3, similar clusters can be detected.

6. Final discussion and related applications

Even though the above findings are only an illustration of the fact that social interactions (based on computational analysis) can be measured, visualised, and explored it is equally obvious that we need to rethink pedagogical theories in order to incorporate the various roles that students can play.

From Figure 3 it is obvious that interesting ideas are conceived by a limited number of "originators" and propagated through the social network. There are several questions that emanate from this observation:

- Does the network pick up and propagate all the relevant/important ideas? If not, how can we make sure that bright ideas are not lost?
- What is the importance of the "transmitter" student? How do we stimulate the students to play the role of transmitter? What facilities are needed in the learning environment?
- What is the role and effect of the popular "high impact" students on overall learning? How can we stimulate the "net importers" to contribute more (increase their impact)? Is it necessary for net importers to gain more impact?
- Do cofactors such as gender and prior knowledge play a role in stimulating the dissemination of ideas? How can the learning environment support this?

Another important conclusion is that many students are located on (rather than inside) the ellipse because they do not heavily rely on the same source (or have a consistent impact on particular students). Therefore their social relationships are greatly determined by random factors that are beyond their control. Again, several questions can be asked about this:

- Is there any difference between students that are located on the ellipse rather than inside it? Does this depend on gender, prior knowledge, age, etc... ?
- Are the learning outcomes different on or inside the ellipse?
- Are any properties of the sociomatrix related to learning outcomes? In other words, is there an ideal social structure or are there certain properties about the social interactions that make a difference in terms of non-rote learning?
- How is the pedagogical paradigm of social constructivism related to the degree of randomness of social interactions in the sociomatrix?

As is often the case, the introduction of a new educational technology solves few problems and raises an abundance of new questions and challenges – Reproducible Computing is no exception to this rule. On the other hand, the quality and quantity of the data that is now available allows educational researchers to investigate the relationships between social networking structure, the role of students therein, and the learning outcomes that result thereof.

Acknowledgements
The Compendium Platform is funded by the OOF 2007/13 project of the K.U.Leuven Association.
I would like to thank Ed van Stee for his useful comments and suggestions.

References

Conole, G., Dyke, M., Oliver, M., and Seale, J. (2004), Mapping pedagogy and tools for effective learning design, Computers & Education 43

de Leeuw, Jan (2001), "Reproducible Research: the Bottom Line", *Department of Statistics, UCLA. Department of Statistics Papers*, nr. 2001031101, http://repositories.cdlib.org/uclastat/papers/2001031101

Dennen, Vanessa Paz (2008), Looking for evidence of learning: Assessment and analysis methods for online discourse, Computers in Human Behavior 24, 205–219

Donoho, D. L., Huo, X. (2004), "BeamLab and Reproducible Research", *International Journal of Wavelets, Multiresolution and Information Processing*

Eggen, P., and Kauchak, D. (2001). *Educational Psychology: Windows on Classrooms (5th ed.)*, Upper Saddle River, NJ: Prentice Hall.

Gentleman, R. (2005), "Applying Reproducible Research in Scientific Discovery", BioSilico, http://gentleman.fhcrc.org/Fld-talks/RGRepRes.pdf

Green, P. J. (2003), "Diversities of gifts, but the same spirit", *The Statistician*, p. 423–438

Khan, Javed I., and Shaikh, Sajid S. (2008), Computing in social networks with relationship algebra, Journal of Network and Computer Applications 31, 862–878

Koenker, R., Zeileis A. (2007), "Reproducible Econometric Research (A Critical Review of the State of the Art)", Department of Statistics and Mathematics Wirtschaftsuniversität Wien, Research Report Series, Report 60

Leisch, F. (2003), "Sweave and beyond: Computations on text documents", *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, ISSN 1609-395X

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

Moody, James (2001), Peer influence groups: identifying dense clusters in large networks, Social Networks 23, 261–283

Peng, R. D., Francesca Dominici, and Scott L. Zeger (2006), "Reproducible Epidemiologic Research", *American Journal of Epidemiology*

Schwab, M., Karrenbach, N. and Claerbout, J. (2000), "Making scientific computations reproducible", *Computing in Science & Engineering*, 2 (6), pp. 61-67

Smith, Erick (1999), "Social Constructivism, Individual Constructivism and the Role of Computers in Mathematics Education", *Journal of mathematical behavior*, Volume 17, Number 4

Von Glasersfeld, E. (1987), “Learning as a Constructive Activity”, in *Problems of Representation in the Teaching and Learning of Mathematics*, Hillsdale, NJ: Lawrence Erlbaum Associates, 3-17.

Wessa, P. (2008a), “A framework for statistical software development, maintenance, and publishing within an open-access business model”, *Computational Statistics*, original publication is available at www.springerlink.com (DOI 10.1007/s00180-008-0107-y)

Wessa, P. (2008b), Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing, Proceedings of the 3rd International Conference on Virtual Learning

Wessa, P. (2009a), Reproducible Computing: a new Technology for Statistics Education and Educational Research, IAENG Transactions on Engineering Technologies, American Institute of Physics, Eds: Rieger, Burghard, Amouzegar, Mahyar A., and Ao, Sio-Iong

Wessa, P., and van Stee E. (2009b), Role and Effect of Reproducible Computing Technology in Statistics Learning, Proceedings of Frontiers in Science Education Research (an International Conference on Science and Mathematics Education Research)

Software:

Csardi, Gabor, and Nepusz, Tamas (2006), The igraph software package for complex network research, InterJournal Complex Systems, pp. 1695

R Development Core Team (2008), R: A Language and Environment for Statistical Computing, ISBN 3-900051-07-0, URL http://www.r-project.org/

Wessa, P., and van Stee, E. (2009), Statistical Computations Archive, K.U. Leuven Association, Belgium, URL http://www.freestatistics.org/

Wessa, P. (2009), Free Statistics Software, Office for Research Development and Education, version 1.1.23-r3, URL http://www.wessa.net/