File 20101022.0554: Weekly activity report 0159:

weekly activity report 159 (loughry)
Joe Loughry
Sent: 22 October 2010 05:54
To: Niki Trigoni; Andrew Martin; Joanna Ashbourn
Cc: otaschner@aol.com; anniecruz13@gmail.com; andrea@hpwtdogmom.org;
chip.w.auten@lmco.com; edloughry@aol.com; diane@dldrncs.com; Joe Loughry;
mmcauliffesl@comcast.net; tom.a.marso@lmco.com

Weekly activity report no. 20101021.1007 (GMT-7) sequence no. 0159, week 2 MT

I submitted confirmation forms to Dr Martin for his signature last
Sunday night. I promised to deliver my confirmation report to him soon.
Written work is due to Julie Sheppard next week; she replied to my email
saying that she would shepherd the forms over to my college for their
official stamp. I leveraged Reading Group this week into a chance to
have a discussion of important issues in my confirmation report with Dr
Flchais, one of my assessors; see below for details.

On Wednesday I was asked to be a technical reviewer for an Army proposal
team bidding on a certification and accreditation contract. The deadline
was short and they needed help right away. I assisted them with updating
their proposal to include the new process for CDS accreditations under
DIACAP using NIST SP 800-53 security controls. In return, I got to
see their whole plan---a blueprint for running a large C&A assembly
line operation. The proposal, number S3R-0354, was submitted on time
to the government yesterday. The proposal was better written than most.

My other project, probabilistic redaction, had its funding turned back
on by the Air Force customer on Monday, ending the stop-work order that
suspended writing of the Quarterly Progress Report and FY 2010 Summary
Progress Report last week. I have to get those reports uploaded as soon
as possible now. To that end, I met with engineer Tom Marso for a demo
of the tool. We discussed implementation details, plans for the next 12
months, and some exciting new capabilities. I will finish my confirmation
report and send it to Dr Martin before I work on the progress reports
for Rome Lab.

The call for papers for the 10th Workshop on the Economics of Information
Security (WEIS 2011) came out this week. The location is GMU, 14--15
June 2011. Submissions are due 28th February. Several people have
told me I should publish my risk market solution at this conference.
The timing is not great, but I will submit to this conference.

The COMLAB-CS-2010 Programme Committee met Wednesday at 10:00am
Oxford time for two and a half hours. The committee ranked all
submissions and selected 16 papers for the programme based on about
95 reviews. We hashed out a timetable and defined and selected four
paper tracks. Next steps are to write the rejection letter, to write
a general-programme-committee-comments report on the overall quality of
submissions with suggestions to authors, and figure out the statistics.

UCDMO has still not published an update to the baseline beyond version
3.5 dated September, 2010. The RM 5.0 developer continues to work on
the baseline patch; internal testers have discovered a few new bugs that
will also be fixed, but nothing serious that affects accreditation. The
developer currently has permission to field TSABI systems but not SABI;
that will proceed after next week's CDTAB deliberation on the next batch
of TORAs, and after DSAWG grants approval for the first SABI (pre-baseline
certification) installation and accreditation. The developer has ODNI's

assurance that RM 5.0 is on the UCDMO baseline; they are merely waiting for the government bureaucracy to finish processing the paperwork.

A press release was being written this week to take some of the wind out of TMAN's sails. TMAN (a competing CDS made by another division of Lockheed Martin) issued a press release last week announcing that they are now on the UCDMO baseline. The developer shared some statistics with me about the relative number and diversity of installations of TMAN vs RM. I was asked to write the new press release, which I did. (I do not believe there is a conflict of interest here.) The COTR and programme office finance people are at the Deer Creek facility this week for business meetings, and are expected to approve the press release. It will be interesting to see how close the final product is to what I wrote.

For the Security Reading Group this week, I introduced a paper from 2006 by Gelman and Stern called 'The Difference Between ''Significant'' and ''Not Significant'' is not Itself Statistically Significant' (The American Statistician 60(4), pp. 328--331). I suggested it in the hope that this paper would provoke a good discussion of experimentation and statistical significance in software engineering and security research.

In the room were John Lyle, Dr Ivan Flchais, Cornelius, Ronald, Shamal, Anbang and myself on video telecon. I began by showing a graph from the Cosmic Background Explorer (COBE) satellite that in 1992 measured the 2.725 K black-body radiation observed over the whole sky with its Far Infra-Red Absolute Spectrometer (FIRAS), a liquid helium-cooled sensor. [Actually, the graph never showed up because screen sharing did not work in Skype, but I emailed it out afterwards.] The graph is interesting because the error bars are 400 sigma wide! Along the entire curve, the difference between predicted and actual values is so small as to be invisible. This is the most precise and accurate measurement of a real system that has ever been done (Barrow, 2006).

So why, I asked, are error bars almost never shown on measurements in software engineering research? Part of the reason, we are told, is because experiments in software engineering are expensive. That brings up an interesting point: the LHC is an expensive experiment too. It cost more to construct than even a large software project. But the LHC, as an experimental device, is probably used to perform millions of experiments over the lifetime of the equipment. I wonder what the fully amortised cost of one experiment is? CERN issues a lot of reports; we could probably calculate it from their financials and their annual reports. Looked at that way, a software project costing 1MM that can only be used once for the purpose of experimentation starts to look a lot more expensive than even specialised scientific apparatus. Someone should do the sums and write a paper about this.

To get the discussion started, I offered the following definition from Krantz (1999): 'an effect is a change in the central tendency of a distribution of measurements, with identical measurements subject to unexplained error.' I stated, for the purpose of argument, that in software engineering a good outcome is often considered to be arriving within 20 percent of your predicted cost/schedule/functionality 80 percent of the time. Ivan immediately pointed out that those are all process metrics. The difference between what we can measure in software engineering and what civil engineers can measure is the difference between taking a sick patient's temperature and knowing that they have bacterial pneumonia. The latter information directly suggests a course of treatment. Fever is a symptom, not the problem itself, just as cracks in a dam are indicative of insufficiently strong materials, or unexpected

stress leading to strain. (I realise that this analogy is shaky.) Mechanical engineers have a wealth of strength-of-materials knowledge to build on. Software engineers seemingly are always implementing new bespoke materials, so they have no already-known performance data. One problem is that, as software writers, we often measure the wrong things. That may be because we do not know what we ought to be measuring, or because the thing we want to measure is impossible to measure, or sometimes because the thing that actually gets measured is the easiest thing around to measure. 'We should always try to measure the property we wish to know directly, rather than measuring another property from which the property we wish to know can be inferred' (Kletz, 1999, p.89). That way lies industrial accidents.

From the paper by Krantz (1999) referenced in Gelman and Stern (2006), there are three schools of thought in psychology when it comes to statistics. Freudian psychologists believe that nothing happens by chance, so to them every event is significant. By contrast, the conventional way that statistics are used in psychological research (which is after all, as John pointed out, mostly applied statistics) is to employ statistical significance tests to distinguish real effects from happenstance. There is a third school of thought in psychology which advocates designing clever experiments in such a way as to maximise the signal to noise ratio so that the effect is obvious. Ivan noted that the last option only front-loads the analysis problem. I countered that it puts the burden on the originator, thereby removing the necessity for having a sophisticated understanding of statistics [what this paper is about] from a much larger number of readers. If that is not such a good end in itself, at least it removes a large number of potential sources of error.

John suggested looking a the provenance community; they are all about reliable data. I pointed out a long quotation from a book by Melton (1962) in which the editors of journals used to replicate results themselves before they would publish a paper; that seems not to be done any more. The standard of evidence, said Dr Flchais, should be to publish enough information to allow others to replicate it. After that others can publish three more papers showing that they were able to replicate your results, or not. Cornelius mentioned that it is possible to choose your statistics in such a way as to get any statistical result you need; this provoked a discussion of lying with statistics.

Ivan: at least in SecHCI, the standards for statistical rigour are non-existent. What, then, is the most meaningful way to compare two pieces of software? Shamal posed a question: to baseline the kind of problem we find in security research, what is the background radiation of the internet? Is it changing over time? Dr Flchais said he would love to see his experimental design. The notion of statistical significance, said Dr Flchais, is a social construct. It has be be within a finite budget; it cannot be arbitrarily high. Ninety-five percent is the accepted level of confidence, so we have no option---we must aim for that as the standard for publishing in this discipline.

We examined Figure 2 in the paper and discussed the presentation of evidence there. Error bars, everyone agreed, are clearly a superior way of presenting information, better than the obfuscatory method used by Blackman et al. (1988). 'If you want high confidence values, and frankly you should to have good science, then you have to choose the areas you measure.' Dr Flchais advised that he would not go below 95 percent.

The point of this paper is to correct a misunderstanding of the meaning of statistical significance by people who use statistics but

are not statisticians. ‘Even large changes in significance levels can correspond to small, nonsignificant changes in the underlying quantities’ (Gelman and Stern, 2006, p. 328). This is different from the more common error of misunderstanding what confidence intervals actually mean, and is completely separate from the question of what significance level is acceptable. The authors present three examples, two drawn from published studies, demonstrating that the authors of those studies claimed that differences in statistical significance were themselves significant, when examination of the underlying data does not in fact support the extended assertion. The relation between the effect size and the standard error must also be taken into account. ‘One way of interpreting lack of statistical significance is that further information might change one’s decision recommendations’ [ibid] and ‘...one should look at the statistical significance of the difference rather than the difference between their significance levels’ [loc. cit.]. These are the key observations of the paper.

I posed a philosophical question to the group: should you publish results that say only ‘we were not able to achieve a 95 percent confidence interval, but we can calculate what we did get, and here it is’? Dr Flchais opined that it is always valuable to know what methods are good and which are bad. He then put me on the spot, asking if the question relates to my confirmation report. I hesitated, but then forged ahead recklessly and replied that that a few months ago I would have answered affirmatively (and in fact that close to what I said publicly in the question-and-answer period after my paper presentation in France). But since then, I have made a conscious effort to improve the methods of my thesis by moving in the direction of a numerical simulation whose parameters I control more than I could control the vicissitudes of an ethnomethodological study. My new study, based on a simple but complete abstract model, grounded by a proof linking it to an old and well-trusted economic theory, implemented through a physical analogue suited to numerical integration, and finally tied to mathematics that are again solvable to arbitrary precision (although not admittedly having a closed form solution!) is my attempt to stay on the straight and narrow path with respect to statistical significance. I can now perform experiments on my model to a 95 percent confidence interval. My abstract model is general enough to describe all the accreditor, developer and certifier behaviour I have observed in two case studies. That is the real direction in which I have made progress.

One of the things this paper says is, if you have a statistical measure that does not reach a 95 percent confidence level, then that tells you something. It tells you that you are not measuring the right things, or you are not using the right tools.

Dr Flchais then asked point blank if I am doing what Kletz teaches not to do. How do I know that I am measuring the value of interest directly, and not some hopefully correlated value in my physical analogue? I replied that my argument lies in the apparent commonality of behaviour seen between the oscillation of risks under multiple bids in the model and the sequence of events observed in two different case studies. I can also report that others I have discussed the model with have told me they feel it is convincing.

Last thing: maybe measuring the only thing you can measure is still a good thing. The discussion moved to the Common Criteria, and whether CC evaluations are mutually comparable. I stipulated that they are not, even when evaluated to the same Protection Profile and EAL, because of the evaluated configuration problem. In addition, the CC itself has evolved radically from version 2.1 to 2.3 to 3.0 to 3.1. The CC

sponsoring organisations are actively trying to make it better---and
they are---but it makes historical comparisons difficult.

I have been talking with Dr Flchais about comparing CC evaluated
products with the CVE again. It sounds like he wants to write a paper
on it.

My current task list (in priority order, most urgent first; work on
tasks in this order):

1. Confirmation report and written work are due to Dr Martin. I have
enough information to write the report now.

2. U.S. Air Force Rome Laboratory quarterly progress report for
probabilistic redaction. After that, I have to combine the last four
quarterly reports to write the FY 2010 summary report.

3. MATLAB work is on hold until I get those three reports and the written
work for confirmation submitted.

4. Finish the Pennock and Wellman (2004) tutorial on uncertainty markets
(still deferred from last week).

5. Follow up with Dennis Bowden and Paul Ozura regarding their recent
emails.

6. Get a date set for telecon with Patti Spicer, Charles Nightingale
and Hal Forsberg at CSC.

7. Meet with Colin regarding audio recording of next OUSS lecture.

8. Implement acid test as an instrumentation process in Simulink.
Figure out how to implement a time-varying spring constant (alarm
clock mechanism) for options. Options either expire or they come
due; need two alarm clocks, not just one.

9. Small tasks: update first case study chart with audience suggestions
from VALID 2010 conference; draw fault-tree diagrams for R-prime,
R-double-prime and S-star; draw up organisation charts for R, R-prime,
S-star, R-double-prime, N, L and G; update documentation of the
current set of anonymisation codes.

10. Crosstalk article: immediately after writing confirmation report,
write the interpretation of the first case study in terms of accreditor
behaviour incentives; write a preliminary overview of second case
study based on final reports from NSA I173 and I733, DNI CAT, ST&E,
POA&M Validation Report, and CDTAB. I have a much better idea of how
to write this now.

Joe Loughry
Doctoral student in the Computing Laboratory,
St Cross College, Oxford

End of WAR 0159.

References