Sociology of large scientific collaborations

Y. P. Viyogi

Recently, during the interview for one of the most prestigious fellowship awards, one of our young colleagues was asked the question. 'How do you feel about being one of the more than 500 authors in any scientific article? How can you be proud of this?' The interview committee consisted of some of the most reputed scientists of our country from different fields, except of course experimental high energy physics, the field in which I work. I do not know whether the reply given by our young colleague was found satisfactory by the committee members, but this certainly compelled me to bring some facts to the notice of the Indian scientific community. Such questions have been asked to almost all of us by fellow scientists working in other areas and whereas some of them have developed appreciation for this social abnormality, most of the others still cannot comprehend how for one physics problem, which is normally to be addressed by one research student either singly or maybe along with his supervisor, there are more than 500 authors. How did those authors contribute to the particular work in question? And most important of all, how does one evaluate the contribution of any individual worker to that published work? Non-appreciation of this aspect has also led to some of our bright young colleagues being denied jobs and faculty positions in many institutions in the country.

I shall try to formulate a method for the evaluation of individual contributions in such large-authorship scientific articles. But before that, let me give a simple example and ask a question. Every decade we conduct a census where huge data are collected for the population of the entire nation. The job of data collection involves millions of workers at the grass-roots level, who are also paid some honoraria for the job. The data become the property of the Government and are analysed only by some of the more privileged scientists, statisticians and others interested in population studies. Suppose a grass-roots worker raises his voice: 'I have also contributed to data collection and hence I should be part of the authorship on any article that is going to be published out of the analysis of these data.' All of us unanimously will rubbish his claim, telling him that he did not do a great job just by going to some houses and recording the statements of people according to some set questions, and also that he has been paid for this. In effect we bought his services and so he cannot lay any claim on the data. The data collection was not a collaborative effort.

The experimental high energy physics community works differently. Scientific collaboration is at its best here and every scientist not only works towards collection and analysis of data, but also renders services which are vital to the running of the experiment. He naturally feels proud to be associated with it as his services are part of the collaborative spirit and not a buy-and-sale arrangement. The collaborative experimental programmes have grown steadily over the years and now at the Large Hadron Collider at CERN, Geneva, one finds close to 3000 authors to a scientific article published by the ATLAS collaboration.

Large scientific collaboration has become common not only in experimental high energy physics, but in almost all branches of science. It is increasing in frequency and importance. Bibliometric studies over the past two decades have shown a continuous increase in the number of co-authored papers in every scientific discipline, as well as within and across countries and geographical areas. The study of scientific collaboration has itself become a topic of intensive research in information science, psychology, management science, sociology and philosophy. It may seem weird, but it is true that the physicists involved in large experimental collaborations at CERN, Fermilab, etc. have also become the subject of study among anthropologists.

The emergence of a good scientific collaboration can be illustrated by the Indian National Gamma Array (INGA) example within our country. Till late nineties the nuclear physicists engaged in the study of gamma-ray spectroscopy formed isolated groups using one or two gamma detectors which each group could afford. Large-acceptance and good-efficiency gamma detectors are expensive, each costing close to a crore rupees. The results from these small experiments were becoming insignificant compared to those coming from European and American groups using large detector arrays like EUROGAM and GAMMASPHERE. The community realized the need to come together and pool in their resources and the result was the INGA. It has detectors contributed by VECC, SINP and DAE-UGC Consortium in Kolkata, IUAC in New Delhi and TIFR in Mumbai. INGA has been used in experiments at all the three accelerators in the country: the Cyclotron at Kolkata and the Pelletrons at Delhi and Mumbai, thus taking advantage of the specific beams in these centres. Starting with a modest array of eight detectors pooled from existing stock in those institutions, the collaboration managed to get DST grant to procure 16 more, and the resulting array is now poised to get international competitiveness in the field of gamma-ray spectroscopy.

Another major collaboration just emerging within the country is connected with the study of neutrinos in an underground laboratory. It has on the list more than 100 scientists and engineers from more than 20 institutions. It is expected to grow further in the years to come when the project gets going and the facility becomes a reality.

Let us now turn to the main question of this note: how does one evaluate the individual contribution of scientists working in large collaborations? In publications emanating from small groups, we tend to give weightage to the first author. This also has its limitations. Senior group members are sometimes known to use alphabetic listing of names as a matter of convenience if their names happen to begin with letters like A, B, etc. Another case arises when the number of authors is not too large but only moderate like 10–15. Such cases are now more common in science than either very small or very large numbers. In the case of publications involving only a couple of groups and having 10–15 names, or those involving just a single large group where a faculty has a large number of students working on very similar problems and the group always writes the names of all the co-workers in the publications, evaluation of the contribution of each

co-worker becomes much more difficult. This is a familiar situation for all of us, which I need not discuss here. But let us see what happens in a very large collaboration where the number of authors could be in hundreds or even thousands.

For such a large collaboration, it has its working rules and procedures for the construction and operation of the experiment, data collection, data analysis and publication of the results. The collaboration management has well-formulated procedures for all types of formats of dissemination of the results, from conferences and meeting to the journal articles. The purpose of this note is to demystify the working of such large and incomprehensible collaborations and also to lay down certain criteria for the evaluation of scientific personnel working there.

Due to its mammoth scale of instrumentation, the experimental apparatus represented by big collaborations is first designed to be modular, with most components having the possibility of being built in small sub-structures and tested individually at remote locations. These are then assembled at the experimental site into the final shape and integrated into the system from the point of view of experimental control, data acquisition and archival. The modularity of the apparatus helps in utilizing the local talent of various groups distributed around the world. This also helps retain individualistic ownership or association and the consequent pride of doing something for the collaboration.

In addition to instrumentation, data archiving, retrieval and processing software has also become a job of enormous magnitude such that several groups now devote their full contribution to the experimental collaboration by way of working for various software packages. The magnitude can be easily gauged by considering that every year each of the four experiments at the CERN Large Hadron Collider will record some 4–5 peta-bytes of raw data. The software effort in each experiment has taken almost the same number of years as the hardware effort in assembling the experimental set-up.

Now let me discuss how the large collaborations manage data analysis and dissemination of the results. Let us take the STAR collaboration as an example, although the rules and procedures are similar in other large collaborations in experimental high energy physics, with minor variants in nomenclature and in some finer details. STAR is one of the two major experiments at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory, USA, with about 500 collaborators. It runs for about six months in a year, where data are collected. The collaboration has an elected Spokesperson, a Council (much like our own legislature) having one member from each collaborating institution and a Physics Analysis Coordinator (PAC). In addition, it has several Physics Working Groups (PWGs), based on various specialized sub-topics of study, which normally have two coordinators. Rules and byelaws of the collaboration are made by the Council using democratic practices. During the six-month running period the collaboration has to man the experiment round the clock and not only ensure that the instruments work properly, but also check and ensure the quality of data being recorded by making quality checks on random samples of data on-line. This requires around 8–10 people in each shift. Every collaborating group has its allocated number of shifts to be taken, depending on the manpower strength of the group (or the number of authors) and to remain on the authorship list the groups are required to fulfil their share of shift responsibility. This ensures that all the groups are fully involved in data collection.

A research scholar who wants to analyse the data for the study of a particular physics problem is required to first approach the relevant PWG coordinator with an expression of interest to pursue such a line of research. This is just to ensure that he is not duplicating the research done by other groups within the collaboration. Usually some form of duplication is allowed (and even encouraged) in special cases, but I will not discuss these details here. He then pursues his analysis and keeps presenting the results in the PWG meetings which are held regularly (almost weekly). Here the research scholar is symbolic; it could also be a small group of individuals having common interest in that particular physics problem and also necessary expertise so that they join together in this work. This small group is called the Principal Authors (PAs). This is very much like doing a piece of analysis on the census data by a small group of statisticians.

At a stage where the analysis becomes mature enough for publication, the PAs in question put all the analysis codes and intermediate figures, tables, etc. (much like the rough work done in solving a problem) in the collaboration's designated web area and prepare a draft manuscript of the work done with the target journal in mind. The PAC in consultation with the Spokesperson appoints a God Parent Committee (GPC), consisting of some five to six members, where the PWG coordinator is a member, one of the PAs is also a member and the Chairperson is usually from outside that particular PWG. GPC also has a member for checking the English language in (sentence formation, grammar and punctuation) the manuscript.

The job of GPC is to go into the details of the analysis, even to the level of checking the codes and reproducing the results. GPC usually has frequent meetings with the PAs and the whole exercise is iterative, with each meeting resulting in some refinement to the text, figures and most importantly the physics goals and conclusions brought out. GPC usually takes about two months to complete its task of vetting the article. It is then released to the entire collaboration for the final checking.

At this last stage, when the work in question is now open to the entire collaboration for opinion, it may take two–three weeks to get it cleared. But in some cases, it may take much longer if there are contentious physics claims which may not satisfy the entire collaboration. The manuscript is submitted to the journal only after the collaboration has agreed to allow it to be submitted for publication. Conflict arbitration is usually done by PAC and the Spokesperson. One may imagine that even if 10% of the collaborators take active part in these reviews, then the manuscript has been refereed by almost 50 scientists before it is sent to the journal. This ensures the high standards of the work and its scientific claims.

If there are referee comments for revision, etc. then depending on the seriousness, the whole exercise may be repeated. The revision is also the job of the PAs. In the simplest form, the revised manuscript is posted for collaboration review at least for one week and then only resubmitted to the journal.

During the progress of the analysis, which may last a few months to several

years, the PAs are allowed to present the intermediate results (labelled 'preliminary') in conferences and symposia and any other similar forums, including presentation for job interviews. Thus till the results appear in print with all the authors, the PAs retain their individual possession. Of course, in a collaboration all the results, whether preliminary or final, belong to the collaboration and if someone other than the PAs is required to present those results, it is always made available. Noting that it is not always possible for each and every student to make presentation in major international conferences, the collaboration encourages the young members to take advantage of several local (national) symposia and conferences, where he can get a chance to make a presentation of the results of the analysis. The slides of any presentation and write-up for the proceedings of conferences have to be usually vetted by PWG and PAC. Unilateral presentation or even submission of abstracts is not allowed. In the case of students it is also required that he must give a rehearsal presentation at least among the local group members in the presence of the Council representative of that institution who in turn will certify to the collaboration about the general standards of slides and physics claims. Publication of the work in all the conferences is normally with only one name (of the person making the presentation), with the collaboration's name added. This also amounts to almost an individual contribution.

It is thus clear that while the analysis of the data is done by a few people (the PAs), the collaboration as a whole retains the responsibility for the authenticity of the results and the physics conclusions drawn from the data. Hence the entire collaboration claims authorship. There have been instances where a particular person (or persons) may not agree with the analysis and the conclusions drawn. He can then withdraw his name from the authorship list, so he feels satisfied that at least he is not going to be responsible for that particular set of results. The authorship policy in large collaborations has been the subject of intense discussions within the community. Even the IUPAP sub-committee on physics had once appointed a committee to look into various aspects and come up with suggestions that would result in smaller number of authors reflecting significant individual contributions. But no acceptable scheme has yet been found.

We find that while being part of a large authorship article does not necessarily mean that one has contributed significantly, one can always distinguish between the PAs and the rest of the authors. The collaboration does not use the full authorship list for work done in connection with instrumentation or the development of certain analysis methods, etc. These are published with only the names of the workers directly involved and have small number of authors like in any other branch of science. The collaboration also publishes its own 'internal notes', which again have less number of authors and are subjected to refereeing by experts within the collaboration. Such works may not have very wide application, but are important for the progress of the physics programme of the collaboration.

Now that we have seen the details of the working of a large collaboration, we can easily formulate a scheme for the evaluation of young workers in this branch of science. Simply going by the large number of publications should not be any reason for acceptance as a bright candidate. Nor should the large number of authors be any reason for rejecting the claim of any good candidate. Our focus should be on the number of articles in which the worker remained one of the PAs, the number of conference presentations, invited talks, short notes published by the collaboration, other publications with less number of authors, etc. These are the works where the individual contribution is reflected most and the concerned person should also possess detailed knowledge of such works. If these are taken into consideration by the scientific community of our country, I am sure there will be justice done to those young workers involved in large collaborative experiments and slowly but surely the community will come to appreciate the importance of this branch of research.

Some interesting reading on scientific collaborations:

1. Sonnewald, D. H., *Annu. Rev. Inf. Sci. Technol.*, 2007, **41**, 643.
2. Shrum, W., Genuth, J. and Chompalov, I., *Structure of Scientific Collaborations*, The MIT Press, 2007.
3. Newman, M. E. J., *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 404; *Phys. Rev.*, 2001, **E64**, 016131.
4. Merali, Z., *Nature*, 2010, **464**, 482; www.nature.com/news/2010/100324/full/464482a.html

Y. P. Viyogi is in the Variable Energy Cyclotron Centre, Kolkata 700 064, India. e-mail: viyogi@vecccal.ernet.in

Soil responds to climate change: is soil science in India responding to?

Manoj-Kumar

Concerns about the likely impacts of climate change on agriculture and the possible implications for future food security have recently fuelled a plethora of research across the various disciplines of agricultural science. This has now fairly improved our understanding of the climate-change impacts on agricultural productivity in different regions of the world. Increasing concentrations of atmospheric CO_2 and the accompanying rise in the earth's surface air temperature, by virtue of their come-along effects, have been recognized as the two most important climate change-associated factors expected to impact crop productivity across the globe. Since atmospheric CO_2 is the sole source of carbon for plants, variations in its concentration have obvious implications for plant growth[1]. The best growth and yield performances of C_3 crops (e.g. wheat and rice) are observed at around 1000–1200 µmol mol^{-1} CO_2 concentration[2,3], implying that the current atmospheric CO_2 concentration (ca. 385 µmol mol^{-1}) is insufficient to saturate the productivity potential of C_3 crops, and hence, any further increase in atmospheric CO_2 is expected to increase the productivity of these crops[4,5]. Unlike the globally observed positive effects of elevated CO_2, impacts of rising temperature are expected