Discrete Mathematics & Big Data
Summary

Peter J. Cameron
University of St Andrews

EPSRC Symposium: Discrete Mathematics & Big Data
University of St Andrews
17 February 2016

I will give a few thoughts of my own, followed by my take on some of the things we have heard over the course of the symposium. It is my own take, but I make no apology: if I misrepresented you, maybe you should have explained it more clearly!

The interdisciplinary nature of the symposium meant that there were many times when people in the audience asked for more detail from speakers. I have not attempted to record this.

Rather than try to draw out general lessons from the very different talks we heard, I have recorded many details about dealing with big data in various contexts, so you can draw your own conclusions.

I hope to make a version of this summary available later.

The combinatorial explosion

How many semigroups (sets with associative binary operation) of order n, up to isomorphism?

n	2	3	4	5	6	7	8	9
#	5	24	188	1915	28634	1627672	3684030417	105978177936292

These numbers are not just evaluations of a formula: essentially the objects must be generated and counted.

Also, future generations of semigroup theorists might need to check the list to test a conjecture or look for a particular property.

So we have to store this data in accessible form.

There are many areas of discrete mathematics where this issue arises!

Open Research Data: UK Concordat

Here are some extracts from the document, my emphasis in red.

- Research Data are quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, interview or other methods.
- The Concordat applies to all fields of research.
- **Principle #3:** Data must be curated so that they are accessible, discoverable and useable.
- **Principle #8:** Data supporting publications should be accessible by the publication date and should be in a citeable form.
- Data underlying publications should be retained for 10 years from collection, creation or generation of the research results.

Example: The Atlas of Finite Group Representations

This site contains a large amount of data on a collection of almost simple groups. It has information on 716 groups in 5215 permutation or matrix representations, some rather large.
For example, $E_8(5)$ has order about $2 \cdot 10^{173}$ and is generated by two 248×248 matrices over the field GF(5).
Sitting at my desk, running my favourite computer algebra program GAP, I can type

```
    > RequirePackage("atlasrep");
```

and then

```
    > G:=AtlasGroup("E8(5)");
```

and generators of the group are downloaded ready for me to explore. Similarly for the other seven hundred groups in the database.
In my view this is a paradigm for big data in discrete mathematics!

Arieh Iserles

The 19th century was the century of steam, the 20th of the internal combustion engine, or of electricity. Will the 21st be the century of data science? Arieh claimed that the correct answer would be information, the organisation of data and its use to transform our lives.

Data science is not just mathematics: it includes topics such as machine learning, image analysis, network analysis and signal processing, but also such non-mathematical aspects as data fusion and curation, natural language processing, legal and ethical aspects, and social science.

The areas of mathematics most relevant to data science are statistics, computation and optimization, applied and computational harmonic analysis, PDEs, and inverse problems.

As individual mathematicians, we should not flock to data science, but continue to do what we are excited about. But heads of department must be aware of funding opportunities and ensure that experts in data science are appointed.

His model of data science is a bicycle wheel. The experts work at the hub; the rim is the entire university; and the spokes are the communication channels.

(I have met this model before. At the Isaac Newton Institute in Cambridge in 2011, John Stufken used it to describe the relationship between theoretical statisticians and scientists who need statistics: essentially the same example, but John thinks of spokes as people who are comfortable both in the hub and at the rim, who can apply the latest theoretical developments to real problems.)

In short, this is a great opportunity for mathematics!

Igor Rivin

Igor began with the notion that discrete mathematicians generate data out of their heads (in the case of Gauss, the first data scientist), or, nowadays, their computers, rather than from experiment or observation; but the principles are the same. As a case study, he described zeolites, hydrated aluminosilicate minerals which now have many industrial uses, from catalysts in the petrochemical industry to cat litter. The number of naturally occurring zeolites is in the hundreds, but vast numbers can be generated by computer; so many, that generating them and searching for those with interesting properties is infeasible. Can we correlate chemical properties of zeolites with their graph-theoretic properties, and so direct the search?

Igor told us the cautionary tale of Doug Lenat who, in the 1970s, started using computers to generate mathematical concepts, and whose name has now given us the unit for measuring bogusity (the microlenat).

A database of potential zeolites is at `https://www.hypotheticalzeolites.net`

V. Anne Smith

Anne was filling in at quite short notice, and told us about Bayesian network analysis of genetic, neural and ecological data.
(It is not so surprising that genomics is connected with discrete mathematics. Eric Lander, the lead scientist on the Human Genome Project, did his doctorate in Oxford on coding theory.)
Anne reminded us that not all "big data" are equally useful. Many observations on a few variables: good. A few observations on each of many variables: not so good!
A Bayesian network describes non-independence between variables: A and B are non-independent if $\mathbb{P}(A \mid B) \neq \mathbb{P}(A)$, and the influence of B on A is mediated through C if $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$. Bayesian networks are always directed acyclic graphs.

There are algorithms which produce a Bayesian network from a given collection of data. For genetic networks involving mRNA, how much of the network can you reconstruct, and how much is wrong? These questions depend sensitively on the amount of data available. A surprising fact to the audience was that, although mRNA produces protein, the mRNA–protein correlations are quite poor.

The methods work much better for neural data, for example, female zebra finches hearing male song; the reconstructed networks agree well with what is known from anatomical studies.

Patric Östergård

Like most discrete mathematicians, Patric had not thought much about Big Data until this meeting gave him the opportunity to step back and reflect.

The problems he works on form a hierarchy: existence; counting (all, or up to equivalence); classification (a description of all objects); and characterization (understanding the objects).

One achievement was the classification of Steiner triple systems on 19 points (there are 11084874829 of them). They were able to store the compressed data in 39Gb (about three bytes per system), but in this form the data is not searchable. By a 72-bit hash function encoding, they produced a 63Gb version of the data, which they were able to use to show that any system can be reached from any other by cycle switching.

The graph is so large that the computer cannot hold the entire thing. We have what he called a big implicit graph, where if we are at a vertex we can find one (or maybe all) adjacent vertices, and we can test whether two vertices are adjacent. “Think global, act local.”

He then told us about his work with Leonard Soicher on the putative McLaughlin geometry. This is an example where the program crunches a huge amount of data, but the answer is likely to be one bit (he guesses “no”).

Rosemary Bailey

Rosemary reminded us that, for the result of data collection to be useful, it is necessary to think beforehand about how it is done (and this is where design of experiments comes in). Statisticians are trained to think that nearly equal replication is crucial, whereas biologists learn that it is important to compare everything to controls. We were shown an example arising in trials of new seed varieties where the best design (minimizing the sum of variances of estimators of treatment differences) shifts from one of these paradigms to the other by a sequence of steps or phase changes as we change the precise number of varieties being tested and assumptions about blocking.

Several earlier speakers mentioned Laplacian eigenvalues, which are also important in discussion of experimental design, where they give us the efficiency factors. Should we think of a huge experiment as an approximation to a manifold?

The design of experiments for large numbers of varieties with very small average replication is still challenging!

Charo del Genio

Charo was also here at quite short notice.
Suppose that observation or experiment has given us a certain network. (He gave an example from the early days of the spread of the AIDS virus, involving people's sexual contacts.) How special is that network? For example, how typical are parameters such as its average distance, among other networks with similar properties? (For mathematicians, a network is just a graph, possibly directed.)
His approach is to generate random networks sharing the appropriate properties, which can be compared with the network we are actually looking at. What kinds of properties are appropriate? He mentioned earlier work on choosing a random network with specified vertex degree sequence.

His recent work involves the joint degree matrix, which gives us (for each α and β) the number of edges between vertices of degree α and vertices of degree β. He showed us how to check the consistency of the JDM, and if it is consistent, how to choose a random network with this JDM.

The question of how to deal with errors in the observed network (a notorious problem in the case of self-reported sexual contacts), or small failures of the degree sequence or JDM to satisfy the necessary conditions, provoked a lively (but inconclusive) discussion.

Also unclear were the hypothesis testing aspects: is the observed network significantly different from a typical one?

Simon Dobson

The title of Simon's talk was "A complex cocktail of networks and reality". He described modelling the transport systems in London and New York as multiplexes (multilayer networks, pairs of networks linked at certain nodes) made up of streets and metro. The first phase looked just at the topology of the networks, and explored shortest paths (in terms of time, assuming a ratio β between one's speed on the street or in the metro). The data is available from Open Streetmap, but needs cleaning before use.

More recent work involves looking at actual flows, using, for example, Oyster Card data from the London Tube.

These networks are far from random, having grown up with many geographical and human constraints. For example, assortativity – the tendency of nodes of high degree to link to other nodes of high degree – influences the spread of epidemics.

This led to a discussion of the Plague, and why (though it still exists, as do rats and fleas), we have not had an epidemic for a long time.

Simon dreams of a project called "Fake Scotland" which would simulate the growth of Scotland, based on these constraints.

He ended with a quote from Alexander Solzhenitsyn, *First Circle*:

> *Topology! The stratosphere of human thought! In the twenty-fourth century it might possibly be of use to someone . . .*

Manish Parashar

Manish began by describing experiments which produce large quantities of data. The Square Kilometre Array will generate an exabyte of data a day by 2020. All branches of knowledge are becoming data-rich. This raises important problems in management and analysis of data.

But all this pales when compared to the amount of data which can be produced by simulation on the latest generation of supercomputers. As we approach exascale science, a machine will need a dedicated power station just to keep it running. (My colleague's response to this was "No wonder these facilities depend not on the National Science Foundation, but on the Department of Energy . . . ")

But there is a huge problem here. Processing speeds have increased enormously, but speeds of data movement have not kept pace. So it is now impossible for a modern supercomputer to save all the data it produces running a simulation! Some data is described as WORN ("write once, read never").

The solution that Manish and his collaborators are working on involves processing the data either in situ or in transit. Multi-core nodes can have some cores producing the data, and others analysing it or constructing visualisations. Another use of local processing is to compress the data so less has to be moved. A further idea is to send the code to where the data is rather than *vice versa*.

This led to an interesting discussion. Arieh Iserles as devil's advocate proposed the thesis that high-performance computing is the enemy of algorithm development, for various reasons: the architectures of these machines make innovative algorithms difficult to run, and the programmers find it easy to be lazy and use the methods they know.

As a general point, with the growth of simulation science, problems of reproducibility arise. As Manish said, software is now critical for the reproducibility of science.

(I couldn't avoid the feeling that as the amount of data grows, the signal-to-noise ratio plummets.)

Franz Király

Franz departed from his prepared talk on a technical aspect of machine learning in order to address the question "What is data science?"

He began with an uncontroversial definition of data, but when he described the scientific method as

> *observation* → *hypothesis* → *prediction* → *experiment* → *cycle*,

many members of the audience objected. In the end we were not going to solve a problem that has been open for millennia, so Franz was allowed to proceed.

The talk quickly specialised to machine learning, supervised or unsupervised. How to measure quantitatively the goodness of a model? This is only defined relative to the data that is being analysed. Given a measure of goodness, we can compare our model; we should compare it to a random guess, and to state-of-the-art or simple models, to see how we are doing.

Ke Yi

By contrast, Ke gave us a nice technical talk about algorithms for sampling from a dataset. We assume that we can't cope with the totality of data, and we wish to sample efficiently. The oldest such result, on random sampling from a data stream, is reservoir sampling. We wish to maintain a random sample of s elements from a data stream. We must start by choosing the first s. If we have a sample from n items, and the $(n+1)$st arrives, we keep it with probability s/n; if we keep it, we discard an existing item (all equally likely). Simple enough, but Ke pointed out that more than half of the proofs of correctness on the first page of a Google search are deficient. He went on to random sampling from distributed streams, range queries, and from data distributed over many nodes, where we want to reduce the amount of communication required.

Jon McLoone

Jon, from Wolfram, told us how to make data science techniques available to a wider audience. One of the key things is keeping the walls as low as possible: rather than cutting-edge programming, the goal is accessibility. All options must have sensible defaults so that the user can progress without worrying about setting them; and since data might be in any format, the translation should just work.

Everything in Mathematica, be it a number, string, picture, database, or program, is a symbolic object. There is no typing in the language because everything has the same type.

Part of the talk was a practical version of Franz Király's talk. When Franz asked about model evaluation, Jon was able to show him a long list of options hiding behind a submenu for the experienced user.

This is not big data; but if it gets a new generation of people interested in capturing and analysing data and producing and deploying the results, it's good for the subject.

Chris Williams

Chris began wearing the hat of his involvement with the EPSRC-funded Centre for Doctoral Training in Data Science in Edinburgh, and moved on to his involvement with intensive care monitoring in a Glasgow hospital.

Data science lies at the intersection of mathematics and statistics, hacking skills, and substantive expertise. If the first element is lacking, we are in a danger zone!

Data science has to deal with questions of scale, fusion of data sources, structure discovery, trust, and ease of use.

We heard about deep learning, a popular topic with students now, but not the answer to all problems in data science.

Clincal data from patients in intensive care (heartrate, blood pressure, etc.) is monitored, of course. But various artefacts affect the data: some is administered by hospital staff (e.g. taking a blood sample, suctioning the lungs), some is caused by the monitoring equipment (e.g. *damped trace*, blockage in the blood pressure monitoring line). The system they have developed is good at detecting some of these such as damped trace (and so reduces false alarms), but not yet good enough for general use.

Rob Ghrist

Rob talked about the use of algebraic topology (specifically persistent homology) in the analysis of certain datasets.

I don't have space here to give a course on algebraic topology, which he described as "the most useful least used mathematics". Our datapoints are often spatial, and we have edges joining certain points, triangles filling in certain triples, and so on. Over your favourite field, you take vector spaces with bases the points, edges, triangles, . . . , and define boundary maps between them reflecting the incidence structure. The quotient of the kernel of one map by the image of the next is a homology group. So these groups reflect geometric aspects of the data.

For example, H_0 tells about connected components; H_1, about cycles; H_2, about hollows bounded by surfaces.

More important, they are functorial: maps between spaces induce maps between the homology groups. Maps between spaces may be given by, for example, changing the scale of measurement.

Now we can decompose homology into indecomposable components. These are described, as the parameter changes, by "barcodes" indicating the points at which they appear and disappear. Components which persist for a long time probably tell us about more interesting features of the data.
Rob mentioned three applications.

- H_0 measures persistent clustering, applicable to genetics, sports data, etc.
- H_1 measures holes in the coverage of sensor networks, and possibly detects whether an intruder can move around without being detected.
- Higher homology has important recent applications in neural connections. The signatures of the dimensions of persistent homology groups in various ranks can distinguish random connections (as in a fly's olfactory system) from geometric connections (as in a rat's visual cortex).

New ideas with promise include using cohomology or sheaf theory, but this is not the place to describe them.

Robert Wilson

Rob began with context, a description of teaching as

$$\text{“reality”} \to \text{data} \to \text{information} \to \text{knowledge} \to \text{wisdom.}$$

These blend into one another without clear boundaries.
In his own field, group theory, the problem is to get information about groups from data giving generating permutations or matrices for the groups.
The Classification of Finite Simple Groups was probably the major achievement of 20th century mathematics. One of these is the Monster, a “sporadic” simple group of order about 10^{54} whose smallest permutation representation is on about 10^{20} points. A single generator for the group would take about 800 exabytes of storage! Matrices are better, since there is a representation by matrices of order 196882 over the field of two elements; one of these only takes about 5 gigabytes.
But these matrices are highly structured, and using the structure and group properties it was possible in the 1990s to fit the generators (and a program) onto a 1.44MB floppy disc.

More recent work has focussed on the idea that, for example, we can study permutation groups without using actual permutations: only a small amount of data in each permutation is actually required. In this way, the computer algebra system GAP can handle groups with big permutation representations (up to about 10^{18} points, so not quite large enough for the Monster yet).

Rob concluded by referring to the Atlas of Finite Group Representations, which I mentioned in my introduction. The Atlas is designed to be useful by ordinary algebraists with no special knowledge of how it was constructed.

Unfortunately, like all public services these days, its continued existence is under threat …

Summing up

- There is a big difference between data generated by a problem in discrete mathematics and data from observation or simulation in science. We have great expertise in the first; can we transfer it to the second?
- Producers of large amounts of data should be encouraged to process it in situ, since moving data is increasingly expensive and slow compared to processing it.
- However the data is produced, there may be people with modest computers and no data science expertise who need to use it. We should store it in a form to make this straightforward.

- It is good to step back sometimes and think about what we are doing. But it is not necessary to have a definition of our subject in order to do it.
- On of the best features of an interdisciplinary meeting like this one is the contacts we have made with people in very different areas.
- This is a great opportunity for mathematics, statistics and computer science to position themselves at the centre of the university and of the "knowledge economy". We should grasp it!