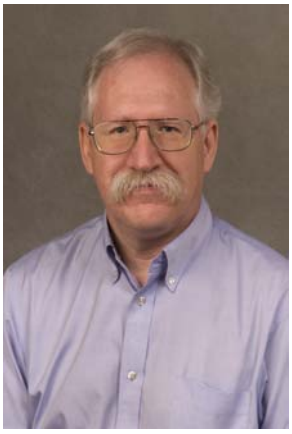
Volume 21, December 2010

A joint newsletter of the Statistical Computing & Statistical Graphics Sections of the American Statistical Association

Statistical COMPUTING & GRAPHICS

A Word from our 2010 Section Chairs



LUKE TIERNEY
COMPUTING

SIMON URBANEK
GRAPHICS

The 2010 JSM saw the presentation of the first Statistical Computing and Graphics Award to Ross Ihaka and Robert Gentleman in recognition for their work in initiating the R Project for Statistical Computing.

Fortunately both were able to receive their awards in person in a very well attended session and present their (very different!) view on future directions for R.

Continued on page 2 . . .

The main purpose of statistical graphics is to display data, models or properties in a way that new insights can be obtained. It is very important that graphics are guided by this principle which is often forgotten in the race for the most fancy or 'cool'-looking displays especially outside of statistical graphics.

Although we should keep our eyes open for new approaches that allow us to display important

Continued on page 2 . . .

Contents of this volume:

Computing Chair
Continued from page 1.

This coming JSM will again have a Data Expo competition. The topic is the Deep-Water Horizon oil spill. For details visit `http://stat-computing.org/dataexpo/`. Data sources are still being collected on the competition web page, and diligent searching may turn up many more, so it is a great opportunity for you to show your skills as a data sleuth.

Thanks to Thomas Lumley for a great program at Vancouver. David Poole has put together a strong invited program for the 2011 JSM at Miami Beach. We also are hoping to offer several exciting CE courses, so you may want to keep an eye out for those. The contributed program of course depends on you: I encourage you to submit contributions and mark them for our section.

This is my final note to you as chair, so I would like tho thank all current and retiring board members for their hard work and welcome the new members and our new chair, Rich Heiberger.

Luke Tierney
University of Iowa

Graphics Chair
Continued from page 1.

properties better or to handle large data, we should always keep in mind that the goal is to represent the underlying data is such a way that we can understand what we see. Only then will graphics serve an analytical purpose.

This may sound familiar to most of our members, but we need to get this message out to other visualization communities as well. Interdisciplinary collaboration is very common and we should take this opportunity to learn from other fields but also share the knowledge we have accumulated in our community over many years. Better graphics even for very specialized applications are high in demand. It is always a pleasure to see excited domain experts when discussing results in graphical form – on paper or interactively.

Part of this sharing at least among statisticians happens at the JSM and we had a very strong program for which I would like to thank Heike Hofmann. Webster West is preparing the next year's program, so please feel free to suggest topic contributed sessions and do not forget to submit abstracts - the JSM website is open for submissions. Also I would like to encourage more graphics contributions to the Student Paper Competition which closes very soon (December 13). Our bi-annual Data Expo 2011 competition is on again - please see the website `http://stat-computing.org/dataexpo/` for details.

Finally, on behalf of the section I would like to thank Rick Wicklin for his excellent service to the section as Secretary/Treasurer and welcome Jay Emerson in this function. I would also like to thank everyone in the Statistical Graphics and Statistical Computing sections as well as the section officers for a great year. Juergen Symanzik will be taking over as the chair of graphics in 2011.

Simon Urbanek
AT&T Research

Highlights from the Joint Statistical Meetings

Stat Computing program

This JSM saw the inaugural Statistical Computing and Graphics award, given to Ross Ihaka and Robert Gentleman for the creation of R. We had a well-attended session including a fairly large contingent from Auckland, where Ross and Robert were working at the time. Among the other invited sessions in Computing was the session on Network models that started off the current SAMSI program on complex networks. Our Topic Contributed sessions included an interesting and diverse set of stu-

dent paper award presentations, congratulations to the students, and thanks to the reviewers. Highlights of the contributed paper sessions included a Tuesday morning session on software and user interfaces. The Joint Computing and Graphics mixer was well attended and we thank the all the companies and individuals who provided door prizes. Finally, the city of Vancouver was itself a highlight of the JSM, a convention location that really is at its best in August.

Thomas Lumley
University of Auckland

Graphical Highlights from JSM 2010

Vancouver was a gorgeous setting for the Joint Statistical Meetings with water and mountains in view from the conference center. The Graphics section had a good turnout of attendees and sessions.

We had two invited sessions this year - the 'Quantified Self: Personal Data Collection, Analysis, and Exploration' had outside speakers, giving us a perspective of collecting data on and around themselves. Seth Roberts from Tsinghua University was sharing his views, his very personal data and got a great discussion going. The other invited session was one that is very close to my heart and research. It introduced two new approaches of getting highly interactive displays into `R`. Simon Urbanek showcased the iplots extreme package `Acinonyx`, Hadley Wickham and Michael Lawrence presented the Qt based packages `qtbase` and `qtpaint`. A few months down the road and more development on all these packages shows more and more clearly that having highly interactive graphics in `R` is shortly to become the rule rather than the exception for everybody!

We had two topic contributed sessions this year. Thanks to Charlotte Wickham and Dawn Woodberger for organizing them! One was looking at last year's data expo - we saw data on millions of flights, and various approaches of how to think about exploring this mass of data. It seemed to be a good venue to have another look at what we only saw briefly in poster form the year before. The other topic contributed sessions linked Bayesian and Graphical Statistics, and provided a theoretical approach to implement user-feedback into a data exploration.

One of the sessions that caused a lot of talk right at the JSM and afterwards, was the session on 'Graphics in clinical trials', which drew so much interest, that people queued in the hallways! In case you were unlucky or missed the session: Jürgen Symanzik pulled all the slides together, and you will be able to find them and session details at `http://stat-computing.org/events/2010-jsm/` .

Overall, one of the highlights of Vancouver's JSM was the session for the inaugural Computing-Graphics Software Award, which went to R & R – Robert Gentlemen and Ross Ihaka – for bringing `R` to us. The award will be ongoing and is used to 'recognize an individual or team for innovation in computing, software, or graphics that has had a great impact on statistical practice or research'. A Revolution Analystics comment summarizes it nicely: "It really can't be overstated how well-deserved this award is.".

Speaking of awards - the section will again sponsor the ASA Data Expo in 2011. It is already on the way, but there is still plenty of time to start working on the data! This year's topic is the effect of the BP oil spill. More details and a sign-up for the competition are available at `http://streaming.stat.iastate.edu/dataexpo/2011/` or just google for 'Data Expo'.

Don't forget to make use of the opportunity to put in topic contributed sessions; it is a great way to get talks on a similar topic into the same session, while at the same time helping to strengthen the section allocating invited sessions in the future. Deadlines for topic contributed sessions are the same as for regular contributed talks.

See you all at Miami Beach next year,

Heike Hoffmann
Iowa State University

JSM 2011 Announcements

Roundtable Discussion Leaders

Dear members of the sections on statistical computing and graphics,

We are looking for people to lead roundtable discussions at JSM 2011 in Miami. Roundtable discussions are small, informal discussions of 10 or fewer people, led by an expert in a particular topic. These roundtables are good ways to meet other folks interested in specific topics in statistical computing, engage in some good discussion, and get a free meal!

Note that leading a roundtable does NOT count towards your one allotted JSM event - i.e. you can lead a roundtable and still be eligible to give another presentation at JSM.
Roundtables in 2010 included:

- Behind the scenes at the Netflix prize
- Facilitating communication about analysis data needs across functional areas
- Effective collaboration of statistical programmers and clinical statisticians
- R's graphical user interface 'R commander' in the intro stats classroom
- R graphics with an Excel front end
- R graphics for EDA

Abstracts for roundtables need to be submitted in January, so if you are interested please contact Chris Volinsky (`volinsky@research.att.com`) and Hadley Wickham (`hadley@rice.edu`).

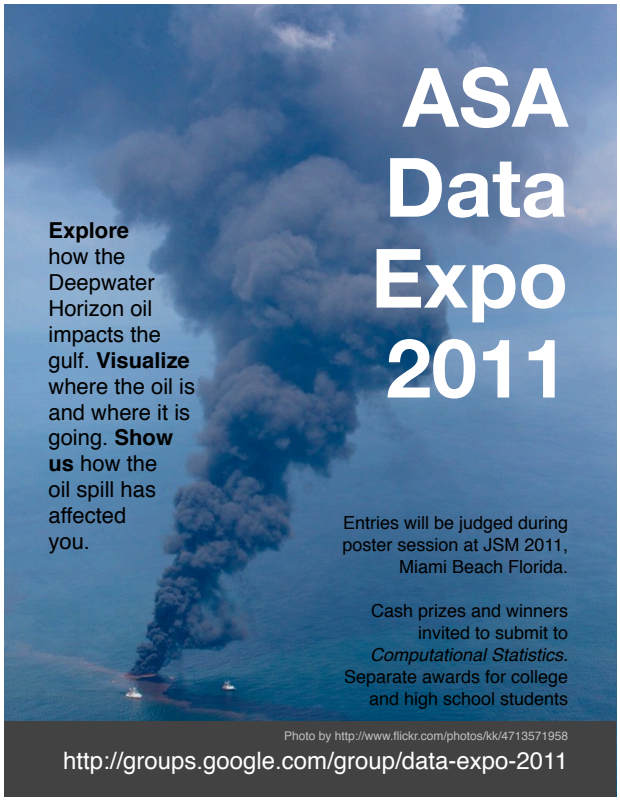
Regards,
Chris Volinsky

Chris Volinsky
AT&T Labs

Data Expo 2011

The ASA Statistical Graphics and Computing sections are pleased to announce Data Expo 2011, a poster competition at the Joint Statistical Meetings in Miami Beach, Florida, Jul 30-Aug 4, 2011. Details of the competition are at:

`http://streaming.stat.iastate.edu/dataexpo/2011/` (or web search "Data Expo 2011".)



This year's data focuses on the government publicly collected data monitoring the effects of the BP oil spill related to April 20 2010 Deepwater Horizon rig explosion.

Entry is open to all. Cash prizes will be awarded. Please send an email expression of interest to `dicook@iastate.edu` if you plan on participating. The first deadline is to submit a poster abstract to the meetings web site by Feb 1, 2011.

Please spread the word as widely as you are willing. The more entrants, the better the competition!

Dianne Cook
Iowa State University

barNest: Illustrating nested summary measures

Jim Lemon and Ofir Levy

Abstract

The challenge of illustrating nested summary measures is introduced and methods for meeting these challenges is outlined. The nested bar plot is suggested as one method that finds support in studies of perception and is similar to other common illustrations. Its major advantage is that most viewers need little or no explanation to understand the relationships illustrated.

Introduction

Descriptive studies of large samples often include a number of categorical measures that define disjunct subsamples, such as males and females or working and unemployed. Summaries of other measures are often of interest, for example the mean ages of males and females or the median income of those working and unemployed. The categories may also be combined to discover whether a comparison within one category is different from that within another. Illustrating such interactions, especially across the different levels of categories, may not be easy.

An example of the problem: women and children first

The classic survival data from the sinking of the Titanic provide a good example of this. The survival rate broken down by accommodation class, sex and age of the passengers and crew is often used to illustrate the unequal outcomes that were once accepted as the normal course of events as well as the gallantry of those who chose to sacrifice their lives for others. Making the relationships between these different effects apparent to the average audience is the challenge that is attempted.

Perceptual considerations

The intended illustration must make both the hierarchical grouping of the summary measures and the comparisons between different levels of grouping obvious. The categories that are combined to make the groups must also be clear to the viewer. The bar plot was chosen as the plot style. One of the principles of gestalt perception is that objects in the visual field will be grouped by their similarities and the groups will be separated by dissimilarities or discontinuities. Thus bar plots in which bars are grouped by their horizonal position provide an easy cue to their relationship in the data set. While it is possible to add an extra level of grouping by stacking bars within a group defined by horizontal position, this makes the relative heights of the bars almost impossible to compare. Additionally, it is difficult to see how this could be extended to further levels of grouping.

A response to the problem: the nested bar plot

The approach taken was to devise a function that would display bars at each level of categorization, making the hierarchical structure obvious by displaying each group of categorized summary measures within the aggregated summary measure that contained them. Thus Figure 1 begins with a bar that is almost the full width of the plot, with each successive set of summary measures plotted within the bar that represents the superordinate category. The three small plots above the main plot show how it is built up. Each level of categorization has a different set of colors, and while the colors are repeated within each level of categorization, it was intended that both the separation between groups and the underlying bars would help to avoid confusion between the groups. The labels beneath the plot were intended to reinforce these distinctions.

Proceeding from the overall proportion surviving, it is clear that the accommodation class had a considerable effect. First class passengers were half again as likely to survive as second class, second

class half again as likely to survive as third class. There was little difference in survival between the third class passengers and the crew. Within the accommodation classes, it is again obvious that children were much more likely to survive than adults, particularly in the first and second classes, in which no children lost their lives. Finally, the innermost sex breakdown shows that females were much more likely to survive in every class and age division apart from the universal survival of children in first and second class mentioned above. Conjectures about the relative gallantry of the males in each class might even be made.

What are the alternatives?

There is an unavoidable compromise in illustrating one summary measure at the expense of others. The nested bar plot attempts to clearly display one summary measure, in this case the proportion of survivors, across the entire hierarchical breakdown. Another approach is to translate the number of survivors and non-survivors in each subset into tiled rectangles, the area of which is proportional to the frequency of each cell. This is known as a mosaic plot, initially attributed to Hartigan and Kleiner [2]. It is possible to construct a mosaic plot with two dimensions of disaggregation on the two axes of a plot and add more levels by nesting (Figure 3). While the mosaic plot accurately portrays the quantitative relationships between the observations, a glance at such a plot of the Titanic data shows the difficulty of visually comparing the proportions across even one level, let alone three.

One commonly suggested alternative to the nested bar plot is the "doubledecker" plot [1]. Figure 2 is an example using the same data with this plot. As the number of children was small in each class and non-existent in the crew, the identifying labels become illegible, even with a wide graphics device. As there is no nesting of the rectangles, the intermediate levels of aggregation are not displayed. All variations on this technique known to the authors suffer from the similar problems. The superbarplot function in the 'UsingR' package uses nested bars to display extreme values and variability, but presents separate subsets of data, such as daily temperature measures.

The nested bar plot departs from the conventional stance that bar plots should be restricted to displaying frequencies. However, the available alternatives are much more difficult, and in some cases impossible, to display in a visual hierarchy.

The programming of barNest

The R function that displays nested bar plots proceeded through a number of stages before producing satisfactory plots. The first consideration was to define and calculate a data object that would contain the summary measures to be plotted. Fortunately, a list containing the summary measures for each level of disaggregation is easy to produce in R.

Displaying the plot requires recursive calls, as the number of category levels typically differ at different levels of categorization. The display is produced by a function that calls itself for each subcategory until there are no further levels of categories to illustrate.

Using the barNest function

Users of 'barNest' will typically have a data frame containing a column of numeric values that are to be grouped by two or more columns of categorical values. The order of grouping is specified in the formula that is the first argument of the function, with the leftmost variable defining the first set of categories. Most of the arguments such as 'main' and 'ylab' will be familiar to those who use other plot functions. The 'trueval' argument is used when the proportion of one value in a catgeorical variable is to be plotted rather than a measure of central tendency in a numeric variable.

The code used to produce Figure 1 is the first example used in the 'barNest' help page in the 'plotrix' package, and recreates the Titanic data set as a data frame. The formula and data set are then passed directly to 'barNest'.exit

In order to define the colors, it is necessary to create a list with one element for each level of categorization. The number of colors in each level should be at least as many as the number of categories. Choosing colors that are easily distinguished may require some thought. Apart from the 'errbars' option, the user may also ask for only the final set of categories to be displayed or not to display the labels below the bars. These bar labels can be passed as a list similar to the bar colors if the labels of the category variables are not suitable. There are a number of fine adjustments like the size

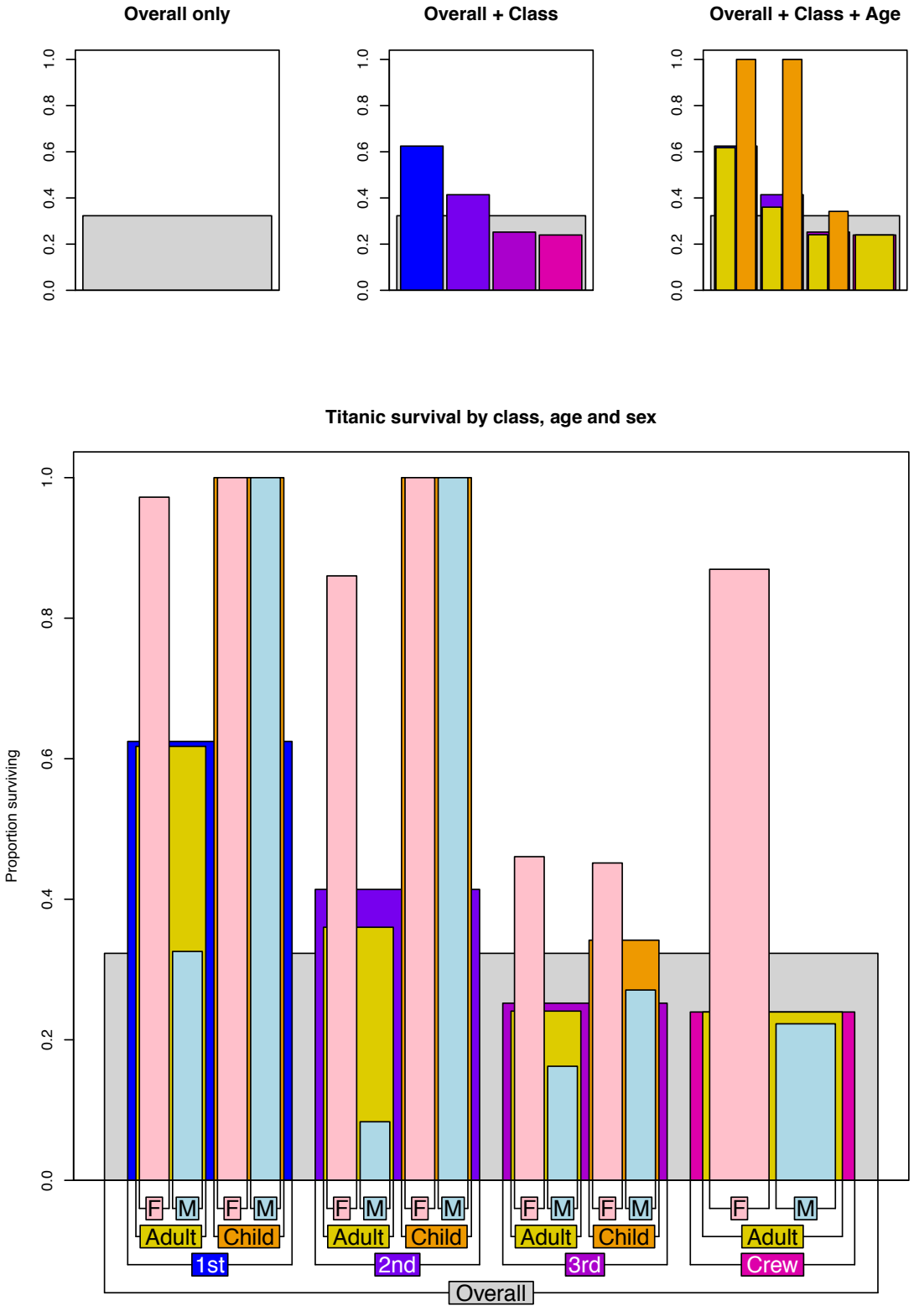


Figure 1: Survival on the Titanic by accommodation class, age and sex (barNest).

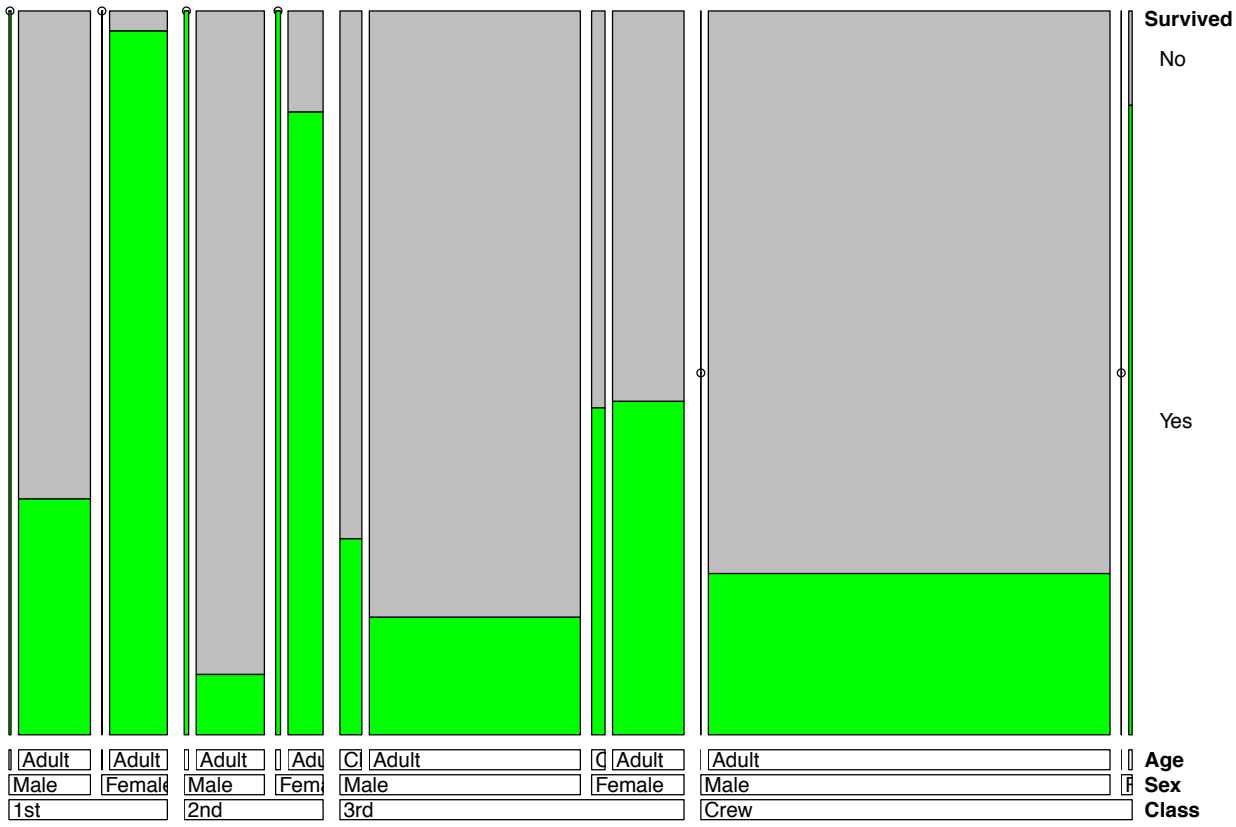


Figure 2: Survival on the Titanic by accommodation class, age and sex (doubledecker).

of the characters in the labels and the proportion to shrink each succeeding set of bars. Detailed instructions for the use of '`barNest`', as well as the function itself, can be found in the '`plotrix`' package.

As a final example of what '`barNest`' can accomplish, we will turn to a lighter set of data concerned with making holes in paper rather than ocean liners. The scores for a year in two matches for a small pistol club will be used to demonstrate an interesting, if sobering, effect of age. The club members were categorized by age as being under 20 years, 20 to 40 years and over 40 years of age. Figure 4 shows the mean scores and 0.25 and 0.975 quantiles broken down by match and age. Only the ultimate breakdown is displayed, but the grouping labels are retained. It is clear that the 20 to 40 year old members have a considerable advantage in both matches. Perhaps just as interesting is that they are also much less variable in their scores. This is not entirely due to the opposing influences of training and age, but is in part a ceiling effect of the possible score of 600 points. However, the ceiling effect is clearly stronger in Open Sport Pistol than in Standard Pistol. Readers who are pistol shooters may be interested in speculating upon this.

Limitations of barNest

Figure 1 displays four levels of summary measures. Even this modest level of complexity may require a wider than normal graphics device to avoid crowding of the labels, although it does not in this instance. The visual complexity of the plot is apparent, yet viewers typically understand the nested structure and grasp the relationships between the values.

The illustration of dispersion available in

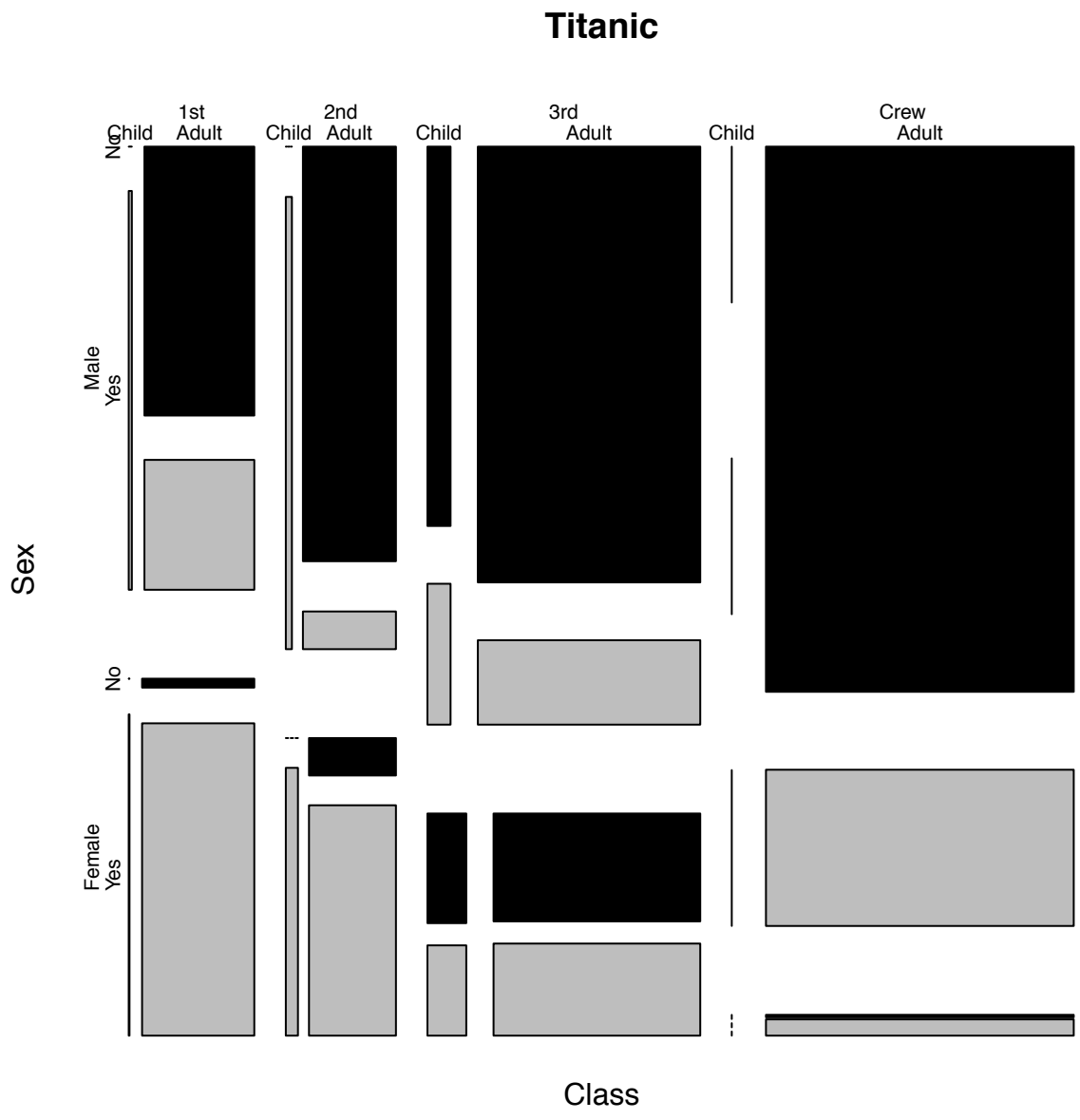


Figure 3: Survival on the Titanic by accommodation class, age and sex (mosaicplot).

'barNest' is easily misused. Setting the 'errbars' argument to TRUE causes the values calculated by the second and possibly third summary functions passed as arguments to be displayed as "error bars". It is left to the user to decide whether these convey valuable, or even correct, information to the viewer. At best, the error bars can give a rough visual indication of the significance of differences in the summary measures, or as in Figure 3, differences in variability that may be of interest.

Summary

The nested bar plot offers the user a method to illustrate summary values that are nested in increasingly complex combinations of categorical variables. Its major advantage is the ease with which an audience can grasp the relationships between hierarchical summary measures with little or no explanation of the illustration. We expect that researchers wanting to illustrate such relationships to

non-specialist audiences, particularly when there is no opportunity to explain the illustration, will find it useful.

Bibliography

[1] M. Friendly. *Mosaic displays for multi-way contingency tables*. *Journal of the American Statistical Association*, (89):190–200, 1994.

[2] J. A. Hartigan and B. Kleiner. *A mosaic of television ratings*. *The American Statistician*, (9):17–23, 1984.

Jim Lemon
bitwrit software
16A Imperial Avenue, Gladesville, NSW 2111
Australia
jim@bitwrit.com.au

Ofir Levy
Department of Zoology
Tel Aviv University, Tel Aviv
Israel
levyofi@gmail.com

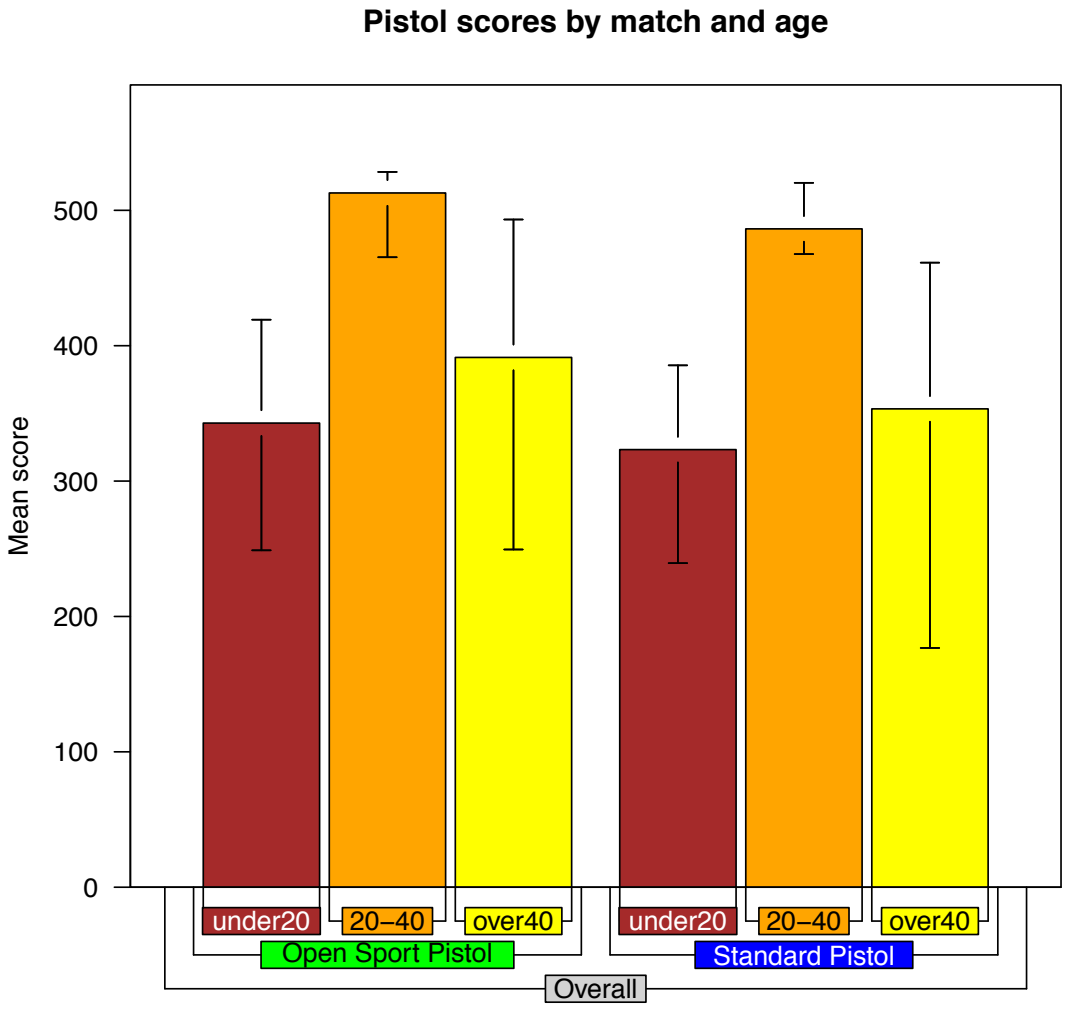


Figure 4: Mean scores and 0.25 and 0.975 quantiles for two matches by age.

You say "graph invariant," I say "test statistic"

Carey E. Priebe, Glen A. Coppersmith and Andrey Rukhin

Introduction: Statistical Inference on Random Graphs

Hypothesis testing on graphs $g \in \mathcal{G}$ has application in areas as diverse as connectome inference (wherein vertices are neurons or brain regions), social network analysis (wherein vertices represent individual actors or organizations), and text processing (wherein vertices represent authors or documents). Graph invariants – functions $T : \mathcal{G} \to \mathbb{R}$ that do not depend on the particular labeling of the vertices – can be used as test statistics on a random graph $G \sim F$ for deciding $H_0 : F \in \mathcal{F}_0$ vs $H_A : F \in \mathcal{F}_A$.

However, even for simple models the exact distribution is unavailable for most invariants. Furthermore, comparative analyses of statistical power at some given Type I error rate for competing invariants, via both Monte Carlo and large sample approximation, demonstrate that simple settings can yield interesting comparative power phenomena.

In particular, two forthcoming articles investigating comparative power of simple invariants in the independent edge setting show that for small graphs (10^3 vertices) the comparative power surface is complicated [1] and limiting behavior may be misleading except for astronomically large graphs [2].

Random Graphs: Simple Models and Simple Invariants

We consider simple graphs (undirected, no weights, no self-loops) on n vertices[1]. Perhaps the simplest of all random graph models is H_0: ER(n, p), wherein all $\binom{n}{2}$ potential edges are independent and identically distributed Bernoulli(p). A simple but interesting and illustrative alternative is given by H_A: $\kappa(n, p, m, q)$, wherein there exists a subset of m vertices and all $\binom{m}{2}$ potential edges amongst this subset are identically distributed Bernoulli(q) while the remaining $\binom{n}{2}$-$\binom{m}{2}$ potential edges are identically distributed Bernoulli(p) with all $\binom{n}{2}$ potential edges independent. We consider p known so that the null is simple; m (and in particular, *which* m) and q are unknown so that the alternative is composite. (These models consist of independent coin flips ... how can this yield interesting (i.e., non-trivial) results? In response to this question we quote Béla Bollobás' comment [3] that a particular J.E. Littlewood paper [4] "is highly recommend to those who think that the binomial distribution is too simple to deserve study.")

Among the simplest of all graph invariants is *size*, the total number of edges in the graph. Under H_0 $size(G)$ is binomial($\binom{n}{2}, p$) while under H_A this invariant is the sum of independent binomials with different success probabilities, limiting Gaussian distributions are available in both cases. Another simple invariant is *maxdegree*, the maximum over all vertices v of the degree $d(v)$. The individual degrees $d(v)$ are binomial($n-1, p$) under H_0 and the sum of independent binomials with different success probabilities under H_A; the collection $\{d(v)\}$ is not independent, but limiting Gumbel distributions are available in both cases.

Many more interesting random graph models – in particular, latent position models – are available for study (see for instance [5] Section 3), as are more elaborate graph invariants – in particular, the number of triangles can dominate *size* and the graph scan statistic can dominate *maxdegree* in terms of statistical power on the inference task (see [1] Figures 13 and 11, respectively). There are significant issues involved in actually computing many candidate invariants on large graphs and in estimation of percentiles for testing, so the trade-offs between better invariants for inference and computational issues demand investigation. Nevertheless, even our simple models and simple invariants yield interesting and illustrative results.

[1] We assume familiarity with basic graph terminology, or see e.g. [1, 2, 3].

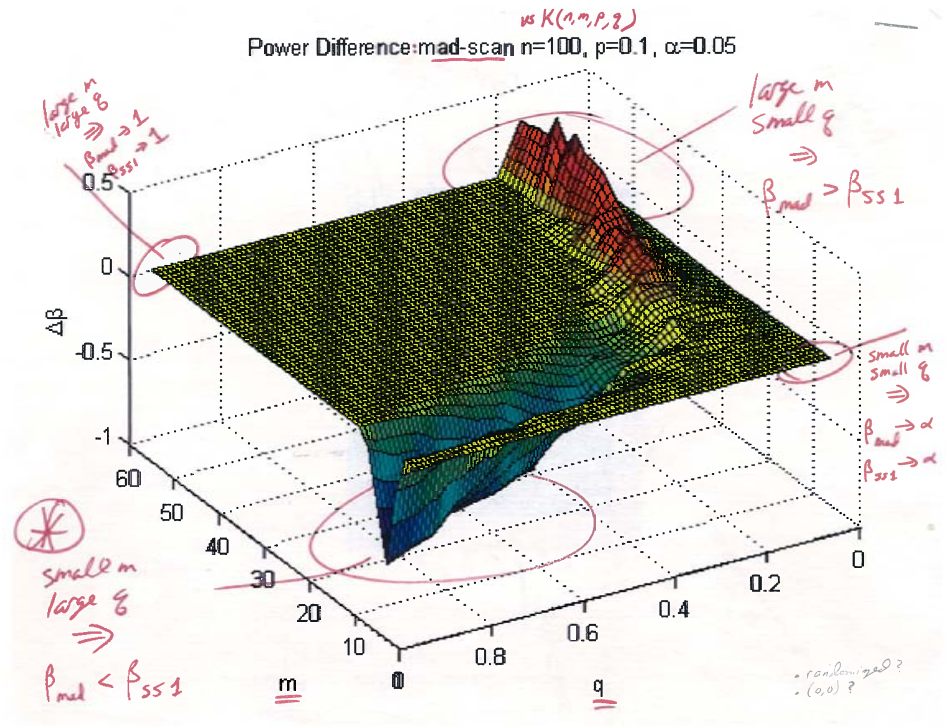


Figure 1: Our first comparative power plot, from 2006. This plot considers the maximum average degree (*mad*) invariant against the graph scan statistic (*scan*) via Monte Carlo. H_0: ER($n = 100, p = 0.1$), H_A: $\kappa(n = 100, p = 0.1, m, q)$, level $\alpha = 0.05$, $\Delta(\beta) \equiv$ power(*mad*)-power(*scan*). We see the phenomenon of interest: both invariants have power $\beta \approx 1$ for large m and q and power $\beta \approx \alpha$ for small m and q, as expected, while for moderate m and q, the comparative behavior of the invariants is quite complicated. In particular, there is a ridge/trough phenomenon running (nonlinearly) from large q small m (where *scan* is superior) to small q large m (where *mad* is superior) which differentiates the invariants. That is, the specific alternative – how many anomalous vertices (m) and by how much are they anomalous (q) – determines the most powerful statistic. This phenomenon is the impetus for on-going research.

Interesting and Illustrative Results

The first plot we made, years ago, which is the impetus for an on-going pursuit of understanding, is shown and described in Figure 1. The ridge/trough phenomenon apparent in the figure begs investigation!

Alas, the maximum average degree is a complicated invariant, both computationally and from a null and alternative distribution perspective, and so we have reason to consider simpler invariants as described above.

The forthcoming *JCGS* paper [1] presents results of a thorough Monte Carlo investigation. For example, Figure 2 is analogous to our original Figure 1 but compares *size* and *maxdegree* at $n = 1000$; the left panel depicts Monte Carlo results, and the right panel depicts the use of asymptotic results to provide approximate distributions analytically. In Figure 2 the Monte Carlo is accurate except for variability due to a finite number of replicates, while the accuracy of the large sample approximations must be verified. For the case depicted ($n = 1000$) the two comparative power surfaces are structurally similar although there are differences beyond just Monte Carlo variation. For smaller graphs ($n = 100$) the Monte Carlo is still accurate but the asymptotic approximations are not; for larger n the Monte Carlo is computationally prohibitive but the asymptotics are accurate. Consider

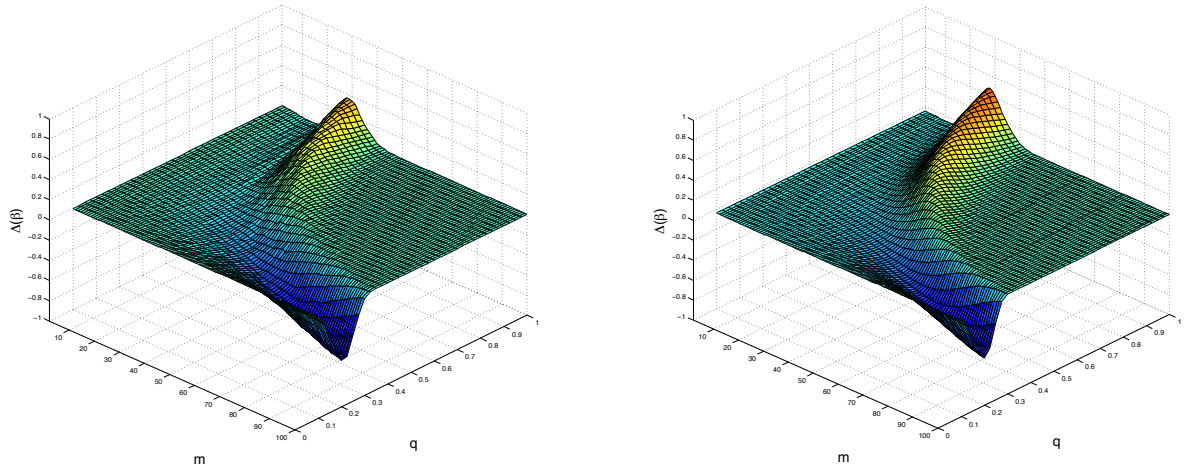


Figure 2: Comparative power plots for *size* vs. *maxdegree* via Monte Carlo (left) and large sample approximation (right) for H_0: ER($n = 1000, p = 0.1$), H_A: $\kappa(n = 1000, p = 0.1, m, q)$, level $\alpha = 0.05$, $\Delta(\beta) \equiv$ power(*maxdegree*)-power(*size*). The ridge/trough phenomenon is apparent. Note that both axes in this Figure have been flipped with respect to Figure 1; with this understanding, one can see that the orientation of the ridge/trough is consistent.

a taxonomy of graph sizes wherein "small" means vertices and edges all fit into memory, "moderate" means vertices fit into memory but not edges, and "large" means even the vertices do not all fit into memory at once. In such a scenario we see that Monte Carlo analysis does not scale well with the number of vertices: for any "large" graph in this taxonomy even the simplest tasks (e.g. performing an edge-count) become computationally challenging.

The forthcoming *JSPI* paper [2] presents theoretical results comparing *size* and *maxdegree*, and shows that when $m = \Theta(\sqrt{n})$ *size* dominates asymptotically but that this domination does not take effect until astronomically large n. This demonstrates that a comparison of test statistics based on limiting power can be misleading for graph inference. Figure 3 illustrates this effect, with $m = \sqrt{n}$.

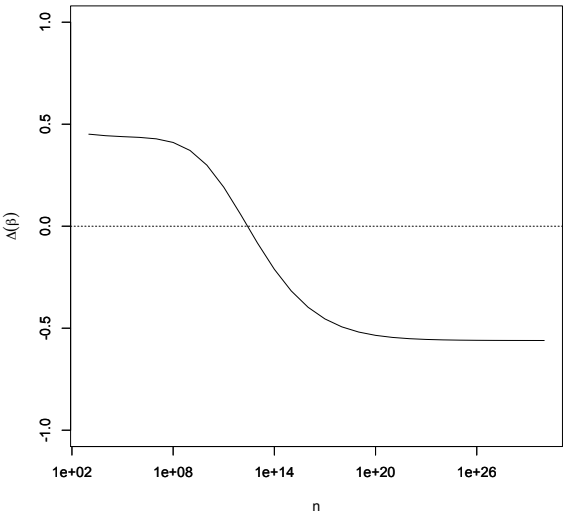


Figure 3: Comparative power plot demonstrating that *size* dominates *maxdegree* asymptotically but that this domination does not take effect until astronomically large n. H_0: ER($n, p = 0.1$), H_A: $\kappa(n, p = 0.1, m = \sqrt{n}, q = 0.9)$, level $\alpha = 0.05$, $\Delta(\beta) \equiv$ power(*maxdegree*)-power(*size*).

Conclusions

The conclusion of this short story is three-fold: (1) even for simple models and simple invariants complex and interesting behavior is evident; (2) much work – theoretical, computational, and experimental – remains to provide results for realistic models and methods; and (3) an interesting plot (such as Figure 1) can provide years of enjoyment!

Bibliography

[1] H. Pao, G. A. Coppersmith, and C. E. Priebe. Statistical inference on random graphs: Comparative power analyses via monte carlo. *Journal of Computational and Graphical Statistics*, forthcoming.

[2] A. Rukhin and C. E. Priebe. A comparative power analysis of the maximum degree and size invariants for random graph inference. *Journal of Statistical Planning and Inference*, forthcoming.

[3] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[4] J. E. Littlewood. On the probability in the tail of a binomial distribution. *Advances in Applied Probability* 1(1):pp. 43–72, 1969.

[5] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning* 2(2):129–233, 2009.

Carey E. Priebe,
Glen A. Coppersmith
Andrey Rukhin
Department of Applied Mathematics and Statistics
Johns Hopkins University,
Baltimore, MD 21218-2682
`cep@jhu.edu`

Computation in Large-Scale Scientific and Internet Data Applications is a Focus of MMDS 2010

Michael W. Mahoney

The 2010 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2010) was held at Stanford University, June 15–18. The goals of MMDS 2010 were (1) to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured scientific and Internet data sets; and (2) to bring together computer scientists, statisticians, applied mathematicians, and data analysis practitioners to promote cross-fertilization of ideas. MMDS 2010 followed on the heels of two previous MMDS workshops. The first, MMDS 2006, addressed the complementary perspectives brought by the numerical linear algebra and theoretical computer science communities to matrix algorithms in modern informatics applications [1]; and the second, MMDS 2008, explored more generally fundamental algorithmic and statistical challenges in modern large-scale data analysis [2].

The MMDS 2010 program drew well over 200 participants, with 40 talks and 13 poster presentations from a wide spectrum of researchers in modern large-scale data analysis. This included both academic researchers as well as a wide spectrum of industrial practitioners. As with the previous meetings, MMDS 2010 generated intense interdisciplinary interest and was extremely successful, clearly indicating the desire among many research communities to begin to distill out and establish the algorithmic and statistical basis for the analysis of complex large-scale data sets, as well as the desire to move increasingly-sophisticated theoretical ideas to the solution of practical problems.

Several Recurring Themes

Several themes—recurring melodies, as one participant later blogged, that played as background music throughout many of the presentations—emerged over the course of the four days of the meeting. One major theme was that many modern data sets of practical interest are better-described by (typically sparse and poorly-structured) graphs or matrices than as dense flat tables. While this may be obvious to some—after all, both graphs and matrices are mathematical structures that provide a "sweep spot" between more descriptive flexibility and better computational tractability—this also poses considerable research and implementational challenges, given the way that databases have historically been constructed and the way that supercomputers have historically been designed. A second major theme was that computations involving massive data are closely tied to hardware considerations in ways that are very different than have been encountered historically in scientific computing and computer science—and this is true both for computations involving a single machine (recall recent developments in multicore computing) and for computations run across many machines (such as in large distributed data centers).

Given that these and other themes were touched upon from many complementary perspectives and that there was a wide range of backgrounds among the participants, MMDS 2010 was organized loosely around six hour-long tutorial presentations.

Large-Scale Informatics: Problems, Methods, and Models

On the first day of the workshop, participants heard two tutorials that addressed computational issues in large-scale data analysis from two very different perspectives. The first was by Peter Norvig of Google, and the second was by John Gilbert of the University of California at Santa Barbara.

Norvig kicked-off the meeting with a tutorial on "Internet-Scale Data Analysis," during which he described the practical problems of running, as well as the enormous potential of having, a data center so massive that "six-sigma" events, like cosmic rays, drunken hunters, blasphemous infidels,

and shark attacks, are legitimate concerns. At this size scale, the data can easily consist of billions to trillions of examples, each of which is described by millions to billions of features. In most data-intensive Internet applications, the peak performance of a machine is less important than the price-performance ratio. Thus, at this size scale, computations are typically performed on clusters of tens or hundreds of thousands of relatively-inexpensive commodity-grade CPUs, carefully organized into hierarchies of servers, racks, and warehouses, with high-speed connections between different machines at different levels of the hierarchy. Given this cluster design, working within a software framework like MapReduce that provides stateless, distributed, and parallel computation has benefits; developing methods to maximize energy efficiency is increasingly-important; and developing software protocols to handle ever-present hardware faults and failures is a must.

Given all of this infrastructure, one can then do impressive things, as large Internet companies such as Google have demonstrated. Norvig surveyed a range of applications such as modelling flu trends with search terms, image analysis for scene completion (removing undesirable parts of an image and filling in the background with pixels taken from a large corpus of other images), and using simple models of text to perform spelling correction. In these and other Web-scale applications, simpler models trained on more data can often beat more complex models trained on less data. This can be surprising for those with experience in small-scale machine learning, where the curse of dimensionality and overfitting the data are paramount issues. In Internet-scale data analysis, though, more data mean different data, and throwing away even rare events can be a bad idea since much Web data consists of individually rare but collectively frequent events.

John Gilbert then provided a complementary perspective in his tutorial "Combinatorial Scientific Computing: Experience and Challenges." Combinatorial Scientific Computing (CSC) is a research area at the interface between scientific computing and algorithmic computer science; and an important goal of CSC is the development, application, and analysis of combinatorial algorithms to enable scientific and engineering computations. As an example, consider so-called fill-reducing matrix factorizations that arise in the solution of sparse linear systems, a workhorse for traditional high-performance scientific computation. "Fill" refers to the introduction of new non-zero entries into a factor, and an important component of sparse matrix solvers is an algorithm that attempts to solve the combinatorial problem of choosing an optimal ordering of the columns and rows of the initial matrix in order to minimize the fill. Similar combinatorial problems arise in scientific problems as diverse as mesh generation, iterative methods, climate modeling, computational biology, and parallel computing. Throughout his tutorial, Gilbert focused on two broad challenges—the challenge of architecture and algorithms, and the challenge of primitives—in applying CSC methods to large-scale data analysis.

The "challenge of architecture and algorithms" refers to the nuts and bolts of getting high-quality implementations to run rapidly on machines, *e.g.*, given architectural constraints imposed by communication and memory hierarchy issues or the existence of multiple processing units on a single chip. As an example of the impact of architecture on even simple computations, consider the ubiquitous three-loop algorithm for multiplying two $n \times n$ matrices, A and B: foreach i, j, k,

$$C(i, j) = A(i, k) * B(k, j).$$

It seems obvious that this algorithm should run in $O(n^3)$ time (and it does in the Random Access Model of computation); but empirical results demonstrate that the actual scaling on real machines of this naïve algorithm for matrix multiplication can be closer to $O(n^5)$. Theoretical results in the Uniform Memory Hierarchy model of computation explain this scaling behavior, and it is only more sophisticated BLAS-3 GEMM and recursive blocked algorithms that take into account memory hierarchy issues that run in $O(n^3)$ time.

The "challenge of primitives" refers to the need to develop algorithmic tools that provide a framework to express concisely a broad scope of computations; that allow programming at the appropriate level of abstraction; and that are applicable over a wide range of platforms, hiding architecture-specific details from the users. Historically, linear algebra has served as the "middleware" of scientific computing. That is, by providing mathematical tools, interactive environments, and high-quality software libraries, it has provided an "impedance match" between the theory

of continuous physical modeling and the practice of high-performance hardware implementations. Although there are deep theoretical connections between linear algebra and graph theory, Gilbert noted that it is not clear yet to what extent these connections can be exploited practically to create an analogous middleware for very large-scale analytics on graphs and other discrete data. Perhaps some of the functionality that is currently being added onto the basic MapReduce framework (and that draws strength from experiences in relational database management or high-performance parallel scientific computing) will serve this role, but this remains to be seen.

New Perspectives on Old Approaches to Networked Data

Although graphs and networks provide a popular way to model large-scale data, their use in statistical data analysis has had a long history. Describing recent developments in a broader historical context was the subject of tutorials by Peter Bickel of the University of California at Berkeley and Sebastiano Vigna of the Università degli Studi di Milano.

In his tutorial on "Statistical Inference for Networks," Bickel described a nonparametric statistical framework for the analysis of clustering structure in unlabeled networks, as well as for parametric network models more generally. As background, recall the basic Erdős-Rényi (ER) random graph model: given n vertices, connect each pair of vertices with probability p. If $p \gg \log(n)/n$, such graphs are "dense" and fairly regular—due to the high-dimensional phenomenon of measure concentration, such graphs are fully-connected; they are expanders (*i.e.*, there do not exist any good cuts, or partitions, of them into two or more pieces); and the empirically-observed degrees are very close to their mean. On the other hand, for the much less well-studied regime $1/n < p < \log(n)/n$, these graphs are very sparse and very irregular—such graphs are not even fully-connected; and when considering just the giant component, there are many small but deep cuts, and empirically-observed degrees can be much larger than their mean. This lack of large-scale regularity is also seen in "power law" generalizations of the basic ER model; it's signatures are seen empirically in a wide range of very large social and information networks; and it renders traditional methods of statistical inference of limited usefulness for these very large real-world networks.

Bickel considered a class of models applicable to both the dense/regular and sparse/irregular regimes, but for which the assumption of statistical exchangeability holds for the nodes. This exchangeability assumption provides a regularity such that any undirected random graph whose vertices are exchangeable can be written as a mixture of "simple" graphs that can be parametrized by a function $h(\cdot, \cdot)$ of two arguments. Popular stochastic blockmodels are examples of parametric models which approximate this class of nonparametric models—the block model with K classes is a simple exchangeable graph model, and block models can be used to approximate a general function h. In this framework, Bickel considered questions of identifiability and consistency; and he showed that, under assumptions such as that the expected degree is sufficiently high, it is possible to recover "ground truth" clusters in this model.

Sebastiano Vigna provided a tutorial on "Spectral Ranking," a general umbrella name for techniques that apply the theory of linear functions, *e.g.*, eigenvalues and eigenvectors, to matrices that do not represent geometric transformations, but instead represent some other kind of relationship between entities. For example, the matrix M may be the adjacency matrix of a graph or network, where the entries of M represent some sort of binary relations between entities. In this case, a common goal is to use this information to obtain a meaningful ranking of the entities; and a common difficulty is that the matrix M may contain "contradictory" information—*e.g.*, i likes j, and j likes k, but i does *not* like k; or i is better than j, j is better than k, but i is *not* better than k.

A variant of this was considered by J.R. Seely who, in an effort to rank children back in 1949, argued that the rank of a child should be defined recursively as the sum of the ranks of the children that like him. In modern terminology, this led to the computation of a dominant *left* eigenvector of M (normalized by row to get a stochastic matrix). A dual variant was considered by T.H. Wei who, in 1952, wanted to rank sports teams and argued that the score of a team should be related to the sum of the scores of the teams it defeated. This led to the computation of a dominant *right* eigenvector of M (with no normalization). Since then, numerous

domain-specific considerations led researchers to propose methods that, in retrospect, are variants of this basic framework. For example, in 1953, L. Katz was interested in whether individual i endorses or votes for individual j, and he argued that the importance of i depends on not just the number of voters, but on the number of the voters' voters, etc., with a suitable attenuation α at each step. Since, if M is a zero/one matrix representing a directed graph, the i, j entry of M^k contains the number of directed paths from i to j, he was led to compute $1\sum_{n=0}^{\infty}\alpha^n M^n = 1(I-\alpha M)^{-1}$. Similarly, in 1965, C.H. Hubbell was interested in a form of clustering used by sociologists known as clique identification. He argued that on can define a status index r by using the recursive equation $r = v + rM$, where v is a "boundary condition" or "initial preference," and this led him to compute $v\sum_{n=0}^{\infty}M^n = v(I-M)^{-1}$.

From this broader perspective, the popular PageRank is the damped spectral ranking of the normalized adjacency matrix of the web graph; the boundary condition is the so-called preference vector; and this vector can be used for various generalizations such as to bias PageRank with respect to a topic or to generate trust scores. Remarkably, although PageRank is one of the most talked-about algorithms ever, there is no reproducible scientific proof that it works for the problem of ranking web pages, there is a large body of empirical evidence that it does not work, and it is likely to be of miniscule importance in today's ranking algorithms. Nevertheless, partly because the basic ideas underlying spectral ranking are so intuitive, there are "gazillions" of small variants that could be (and are still being) introduced regularly in many areas of machine learning and data analysis. Unfortunately, this is often without reproducible scientific justification or careful evaluation of which variants are meaningful or useful.

Matrix Computations—in Data Applications

Challenges and tradeoffs in performing matrix computations in MMDS applications were the subject of the final pair of tutorials—one by Piotr Indyk of the Massachusetts Institute of Technology, and one by Petros Drineas of Rensselaer Polytechnic Institute.

Indyk discussed recent developments in "Sparse Recovery Using Sparse Matrices." This problem arises when the data can be modeled by a vector x that is sparse in some (often unknown) basis; and it has received attention recently in areas such as compressive sensing, data stream computing, and combinatorial group testing. Traditional approaches first capture the entire signal and then process it for compression, transmission, or storage. Alternatively, one can obtain a succinct approximate representation by acquiring a small number of linear measurements of the signal. That is, if x is an n-vector, the representation is Ax, for some $m \times n$ matrix A. Although typically $m \ll n$, the matrix A can be constructed such that one can use a recovery algorithm to obtain a sparse approximation to x. It is often useful (and sometimes crucial) that the measurement matrix A be sparse, in that it contains very few non-zero elements per column. For example, sparsity can be exploited computationally—one can compute the product Ax very quickly if A is sparse. Similarly, in data stream processing, the time needed to update the sketch Ax under an update Δ_i is proportional to the number of non-zero elements in the i-th column of A.

Indyk described tradeoffs that arise when designing recovery schemes to satisfy the tricriterion of short sketches, low algorithmic complexity, and strong recovery guarantees. Randomization has proved to be an important computational resource, and thus a key issue has been to identify properties that hold for very sparse random matrices and also are sufficient to support efficient and accurate recovery algorithms. A key challenge is that, whereas dense random matrices are fairly homogeneous (*e.g.*, since measure concentrates their eigenvalues follow Wigner's semicircle law), very sparse random matrices are much less regular. One can say that a matrix A satisfies the $RIP(p, k, \epsilon)$ property if

$$||x||_p(1-\epsilon) \leq ||Ax||_p \leq ||x||_p$$

holds for any k-sparse vector x. (This generalizes the well-known Restricted Isometry Property from $p = 2$ to general p.) Although very sparse matrices cannot satisfy the $RIP(2, k, \epsilon)$ property, unless k or ϵ is rather large, Indyk showed that the adjacency matrices of constant-degree expander graphs do satisfy this property for $p = 1$ and that several previous algorithms generalize to very sparse matrices if this condition is satisfied.

In his tutorial on "Randomized Algorithms

in Linear Algebra and Large Data Applications," Petros Drineas used his work on DNA single-nucleotide polymorphisms (SNPs) to illustrate the uses of randomized matrix algorithms in data analysis. SNPs are sites in the human genome where a nonnegligible fraction of the population has one allele and a nonnegligible fraction has a second allele. Thus, they are of interest in population genetics and personalized medicine. In addition, they can be naturally represented as a $\{-1, 0, +1\}$ matrix A, where A_{ij} represents whether the i-th individual is homozygous for the major allele, heterozygous, or homozygous for the minor allele.

While some SNP data sets are rather small, data consisting of thousands or more of individuals typed at hundreds of thousands of SNPs are increasingly-common. Size is an issue since even getting off-the-shelf SVD and QR decomposition code to run on dense matrices of size, say, $5000 \times 500,000$ is nontrivial on commodity laptops. The challenge is especially daunting if the computations need to be performed thousands of times in the course of a cross-validation experiment. Perhaps less obvious is the issue of interpretability—even if the data clusters well in the span of the top k "eigenSNPs," these eigenSNPs cannot be assayed in the lab and they cannot be easily thought about. Thus, while eigenvector-based methods for dimensionality reduction are popular among data analysts, the geneticists were more interested in the k actual SNPs that were most important.

Drineas described how to address these two challenges—the "challenge of size" and the "challenge of interpretability"—in a unified manner. He described a randomized approximation algorithm for choosing the best set of exactly k columns from an arbitrary matrix. The key structural insight was to choose columns according to an importance sampling distribution proportional to the diagonal elements of the projection matrix onto the span of the top k right singular vectors. These quantities can be computed exactly by computing a basis for that space, or they can be approximated more rapidly with more sophisticated methods. Importantly for interpretability, these quantities are the diagonal elements of the so-called "hat matrix," and thus they have a natural interpretation in terms of statistical leverage and diagnostic regression analysis. Importantly for size and speed, Hadamard-based random projections approximately uniformize these scores, washing out interesting structure and providing a basis where simple uniform sampling performs well. This has led in recent years to fast high-quality numerical implementations of these and related randomized algorithms.

Conclusions and Future Directions

In addition to these tutorial presentations, MMDS participants heard about and discussed a wide range of theoretical and practical issues having to do with algorithm development and the challenges of working with modern massive data sets. As with previous MMDS meetings, the presentations from all speakers can be found at the conference website, `http://mmds.stanford.edu`; and as with previous MMDS meetings, participant feedback made it clear that there is a lot of interest in MMDS as a developing research area at the interface between computer science, statistics, applied mathematics, and scientific and Internet data applications. So keep an eye out for future MMDSs!

Acknowledgments

I am grateful to the numerous individuals who provided assistance prior to and during MMDS 2010; to my co-organizers Alex Shkolnik, Petros Drineas, Lek-Heng Lim, Gunnar Carlsson; and to each of the speakers, poster presenters, and other participants, without whom MMDS 2010 would not have been such a success.

Michael Mahoney (`mmahoney@cs.stanford.edu`) is in the Department of Mathematics at Stanford University. His research interests include algorithmic and statistical aspects of large-scale data analysis, including randomized algorithms for very large linear algebra problems and graph algorithms for analytics on very large informatics graphs.

Bibliography

[1] G.H. Golub, M.W. Mahoney P. Drineas, and L.-H. Lim, "Bridging the gap between numerical linear algebra, theoretical computer science, and data applications," *SIAM News*, **39**, no. 8, (2006).

[2] M.W. Mahoney, L.-H. Lim, and G.E. Carlsson, "Algorithmic and Statistical Challenges in Modern Large-Scale Data Analysis are the Focus of MMDS 2008," *SIGKDD Explorations*, **10**, no. 2, pp. 57–60, (2008).

Michael W. Mahoney
Dept. of Mathematics
Stanford University
mmahoney@cs.stanford.edu

Section Officers

Statistical Computing Section Officers 2010

Luke Tierney, Chair
luke@stat.iowa.edu
(319) 335-3386

Jose C. Pinheiro, Past Chair
jpinhei1@its.jnj.com
(908) 927 5204

Richard M. Heiberger, Chair-Elect
rmh@temple.edu
(215) 808-1808

Usha S. Govindarajulu, Secretary/Treasurer
usha@alum.bu.edu
(617) 525-1237

Montserrat Fuentes, COMP/COS Representative
fuentes@stat.ncsu.edu
(919) 515-1921

Thomas Lumley, Program Chair
t.lumley@auckland.ac.nz
+64 9 373 7599 ext 83785

David J. Poole, Program Chair-Elect
poole@research.att.com
(973) 360-7337

Barbara A Bailey, Publications Officer
babailey@sciences.sdsu.edu
(619) 594-4170

Jane Lea Harvill, Computing Section Representative
Jane_Harvill@baylor.edu
(254) 710-1517

Donna F. Stroup (see right)
Monica D. Clark (see right)

Nicholas Lewin-Koh, Newsletter Editor
lewin-koh.nicholas@gene.com

Statistical Graphics Section Officers 2010

Simon Urbanek, Chair
urbanek@research.att.com
(973) 360-7056

Antony Unwin, Past-Chair
unwin@math.uni-augsburg.de
+49-821-598-2218

Juergen Symanzik, Chair-Elect
symanzik@math.usu.edu
(435) 797-0696

Rick Wicklin, Secretary/ Treasurer
Rick.Wicklin@sas.com
(919) 531-6629

Peter Craigmile, GRPH COS Rep 08-10
pfc@stat.osu.edu
(614) 688-3634

Mark Greenwood, GRPH COS Rep 10-12
greenwood@math.montana.edu
(406) 994-1962

Heike Hofmann, Program Chair
hofmann@iastate.edu
(515) 294-8948

Webster West, Program Chair-Elect
websterwest@yahoo.com
(803) 351-5087

Donna F. Stroup, Council of Sections
donnafstroup@dataforsolutions.com
(404) 218-0841

Monica D. Clark, ASA Staff Liaison
monica@amstat.org
(703) 684-1221

Martin Theus, Newsletter Editor
martin@theusRus.de

Statistical
COMPUTING & GRAPHICS

The Statistical Computing & Statistical Graphics Newsletter is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding the publication should be addressed to:

Nicholas Lewin-Koh
Editor Statistical Computing Section
lewin-koh.nicholas@gene.com

Martin Theus
Editor Statistical Graphics Section
martin@theusRus.de

All communications regarding ASA membership and the Statistical Computing and Statistical Graphics Section, including change of address, should be sent to

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402 USA
TEL (703) 684-1221
FAX (703) 684-2036
asainfo@amstat.org