NYU Seminars

JHD

2 February 2017 —

Contents

Chapter 1

Hardware-conscious data processing systems by Holger Pirk

3 February 2017.

1.1 Speaker's Abstract

Performance engineering, i.e., the processes of tuning the implementation of an algorithm for a given set of hardware, application and data characteristics, can reduce query response times of data processing systems from minutes to milliseconds – it turns long-running jobs into interactive queries. However, when building such systems, performance is often at odds with other factors such as implementation effort, ease of use and maintainability. Well-designed programming abstractions are essential to allow the creation of systems that are fast, easy to use and maintainable.

In my talk, I demonstrate how existing frameworks for high-performance, data-parallel programming fall short of this goal. I argue that the poor performance of many mainstream data processing systems is due to the lack of an appropriate intermediate abstraction layer, i.e., one that allows the hardware and data-conscious application of state-of-the-art low level optimizations.

To address this problem, I introduce Voodoo, a data parallel intermediate language that is abstract enough to allow effective code generation and optimization but low-level enough to express many common optimizations such as parallelization, loop tiling or memory locality optimizations. I demonstrate how we used Voodoo to build a relational data processing system that outperforms the fastest state-of-the-art in-memory database systems by up to five times. I also demonstrate how Voodoo can be used as a performance engineering framework, allowing the expression of many known optimizations and even enabling

the discovery of entirely new optimizations.

1.2 Preliminaries

Intro This was the first Faculty Hiring talk in the new building.

Speaker at MIT, ex CWI.

1.3 Performance

Example 1 (TPCH Benchmark problem 1) *We have a typical sales query, where 100 seconds is a typical time for a database query against a 10GiB dataset (note not even disc bound!). But this dataset is small enough to fit in main memory these days. But then this would take 37ms to read the data, so there's a real issue. PostGres on a RAM disc actually drops to 96 seconds. CWI's MonetDB [BKM08], specialized for this, is 3.2s. At MIT, with Voodoo our name) we have 0.162 seconds. How did we get there?*

MonetDB has specialised data analytic kernels. . Classic architecture is SQL $\Leftrightarrow$ Logical Plan (Select, Join etc.) $\Leftrightarrow$ Physical Plan (TableScan etc.) $\Leftrightarrow$ DB kernel $\Leftrightarrow$ OS/hardware. Our next step in Voodoo is essentially compiling, in this case from the logical plan (2$\times$Join+Select) to code. So a Join becomes creating (first argument) and probing a hash table. This is good for a 4$\times$ speed-up. This is basically what a smart undergraduate would do.

Then the undergraduate learns about the Performance Engineering Deluge. Many *ad hoc* point solutions,

- Not portable
- Integration into the system is hard
- Interplay of techniques is hard to understand
- Impact of data statistics (locality etc.) is unclear.

Hence few systems implement any techniques, and none implement all. What about Co-processing, SIMD Processing, indices etc. Hence I propose a unified intermediate algebra

fast like C

Algebraic To allow relational algebra optimisation

Portable At least not break when moved, even if optimisation doesn't improve things.

1.4 State of the Art

Each hardware technique has its own "abstraction", e.g. CUDA for GPUs, or compiler intrinsics for SIMD. In his TBB, he has `parallel_deterministic_reduce`, per-partition λs and global λs. Using SIMD, we partition the data into lanes, and end up casting things into vector types etc. Conceptually similar, but code techniques totally different.

1.5 Voodoo

We have taken the SQL $\Leftrightarrow$ Logical Plan from MonetDB.

Fast All tuning decisions explicit.

All operators in Voodoo are parallel: `Project`, `Zip`, `Gather` etc. The new idea is controlled folding. Standard `fold`, but assign partition ids to each datum, then fold by partitions. For example `PartitionSize:=Constant(4)`; then partitioning functions. Hence the difference between multicore and SIMD is whether we use `Divide` or `Modulo` as the partitioning function. Chosen manually, but we are exploring automated tuning. These do compose, so SIMD multicores are supported.

Q GPU/CPU hybrid co-processing?

A Not currently supported, but expressible.

1.6 Voodoo SQL frontend

Hence the graduate version of the query has a multicore `Divide` before the `fold`.
Shows gains across all the TPC benchmark questions: all good.
Extends the query to add a Foreign Key join.

Q Why not implement in LLVM?

A It does a poor job of data optimisation. Hyper does this, but we are significantly faster than Hyper.

Chapter 2

Parallel I/O Performance: Andy Turner

Actually an Archer Webinar on 8 February 2017, which JHD joined remotely, but can't see anywhere else convenient to stash his notes.

This work came out of users saying that this was hard to understand. Typically, people have worked on other performance-related things, and I/O is what's left. What's rare is practical studies, as opposed to benchmarking. Note that there's user control of Lustre striping settings, which gives an added variable.

2.1 Benchmarks

SSF = "Single Shared File". FPP = "File Per Process". Note that SSF is "received wisdom".

IOR is commonly used, especially at procurement. We didn't choose this: it's opaque (many options, but hard to work out what I/O commands are actually being performed). Not a good mapping between its options and what users actually do. IOR is the Linpack of I/O — maximum rather than useful.

Benchio (SSF) A simple FORTRAN program. Only does write performance. `https://github.com/EPCCed/??`.

Benchio_FPP (FPP) We implemented this (same GitHub). Again FORTRAN. Needs to be careful not to be distorted by buffering.

2.2 Machines etc.

Archer In theory 30GiB/s Lustre.

COSMA5 In theory 20GiB/s. GPFS

Jasmin A Panases system. Data not yet really analysed.

MPI Use collective operations (necessary!).

Experiments tried to run at different times of day/week to get a good mix.

2.3 SSF

Lustre: had to use maximum stripe counts: good for a factor of $\times 10$ at best, but quite a wide performance variation (2–15 GiB/s). GPFS was much less variable.

Needs well-written parallel libraries and parallel collective I/O. SSF provides a simple end-product: one file capable of simple re-use.

2.4 FPP

Good performance once over four nodes. Again Archer is much more variable than GPFS. With striping, we had failures due to excessive metadata operations. Hence recommend single striping for FPP.

2.5 Comparison

See slide 20. At low core counts, FPP seems to win. SSF may be taking over at 64 nodes (i.e. #nodes > maximum number of stripes), but even then not by much. FPP looks great for checkpointing, but you need the same decomposition to restart.

Both schemes can achieve 50% of the theoretical peak (open question: why only 50%)?

Q JHD: you mention node counts: were you using all cores on these nodes?

A Yes, all cores were writing, but of course there's aggregation at the node level before the requests leave the node.

Q Striping?

A `lfs setstripe -c -1 <dir>` is the usual way of doing this. There's an API way, but only IOR does this. Note that the usual technique is to create an empty directory then do this, so all new files inherit the striping. The command doesn't restripe existing files.

Chapter 3

PowerShell Empire

3.1 Preliminaries

This is about PowerShell, but about far more than that: PowerShell. Note that this access to a wide range of tools. This is great for sysadmins, but also gives the bad guys a lot of power.

Diagram shows a connection over http/https between Empire server and Empire Agent. The server runs on a Linux box. I believe it's written in Python. The agent is pure PowerShell, running on the victim machine. There are four components

Listeners URLs on the server

Agents The PowerShell that runs on the victim

Stagers how agents get onto the victim. Can be (he shows) a one-line PowerShell command using `-Enc` and an encrypted piece of PowerShell code. The stager depends on the listener.

Modules what you can do once an agent connects to your server. "180 and counting".

Demonstrates issuing commands from the Empire Server to the Agent. Likens this 'server' to the 'Adobe download program', which is small, but does the real download.

Q Can't you just disable PowerShell?

A Many admin tools are GUIs on top of PowerShell, so no?

Q Authorised scripts only?

A Works in the simple case (but only then).

Get a Kali Linux, then

```
cd /opt
git clone https://github.com/PowerShellEmpire/Empire.git
cd Empire/setup
./install.sh
./empire
```

3.1.1 Policies

"Disable PowerShell for end users" — this is possible. Needs AdBlocker or equivalent (DeviceGuard),

3.1.2 Detection

Basically logging, where 'script block logging' can be turned on. This is effective because the agent is permanently checking in with the server. Can turn on transcripts, and get even more.

`psinject` in a module inside PowerShell Empire, which, using DLL injection and "reflection", can inject empire into a running process like Notepad (easy) or a privileged command like `lsass` (harder). Then PowerShell logging goes silent, as the work is being done elsewhere. Defences have improved in the latest PowerShell, but there are modules that download old versions of PowerShell in order to avoid these.

From the system log, one can see events like "Notepad starts Regedit", which is unusual. Note that parameters get logged, so passwords might be captured inadvertently this way. "List DLLs" might also be helpful: compare its results with a healthy system.

3.2 Conclusions: Carbon Black

PowerShell Empire is a classic example of a non-malware attack, described as "living off the land". Hence techniques like application whitelisting are rendered ineffective. Our `AdBlocker` is a start. We regard antivirus etc. as "point-in-time" prevention, whereas our latest product (CB Defense) is "streaming prevention", so we capture events and analyse them.

Example 2 *Firefox calls Flash (normal) calls (PowerShell) dubious, which starts a process and suspends it, at which point we realise there's definitely something dubious.*

Q Why don't I see the performance hit?

A Multicore.

Q Does this use a special VPN?

A No. But it is a cloud-only solution: you need a connection.

Q Does it connect to other CB products?

A Not yet, but things coming this spring.

Q "Long and slow attacks"?

A I don't know how slow the attack has to be to get under our radar.

Q How vulnerable us your agent?

A I've not seen any PowerShell attacks. But in general there are attacks (we have various defences, of course), hence a broad approach to defence is important.

Q Defense versus Response?

A Response can be run on-site. It's more like an observer than Defence, which is a pro-active guard.

Q Mac?

A Defense and Response both do this.

Q Learning? So if you see a lot of A/B/C/D/E/block, how about blocking at A/B/C/D/block?

A What about the A/B/C/D/G valid paths. There's a real false Positive issue here.

Q We couldn't get the advanced Microsoft features to run in early Windows 10.

A Microsoft is leaning a lot harder on manufacturers over hardware.

Chapter 4

13 February 2017 (Princeton)

Ed Felten - West Wing, Veep, or House of Cards? Policy and Technology in the Obama White House JHD attended this remotely, prompted by JJB. There was some difficulty fitting the local audience into the room. Advertisement for a talk "Why Fighting Online Abuse is Good for Free Speech".

4.1 Abstract

Professor Felten recently completed a 20-month tour of duty at the White House, where he served as Deputy U.S. Chief Technology Officer. In this talk he will discuss his experience there, describe ongoing policy challenges, and reflect on the role of technology and academic expertise in policymaking.

4.2 Actual Talk

Thanks Family, and colleagues in White House. "It's a team sport". Also Princeton for allowing it.

May 2015 became Deputy US CTO. Media portrayals of the White House differ. "West Wing" — dedicated and super-competent. There *is* a lot of walking and talking. Mostly werededicated and competent. "Veep" — ego and dysfunction. "Forgetting to bring socks to work" did actually happen, but are not common. What is true is that little mistakes get magnified. "House of Cards" — deception and limitless will to power. I did not see any bloodshed.

I'll talk about serious policy matters, but also detours about life. I was there (Oval) to give the President a detailed briefing on a technical topic. Half of my brain was in awe (JHD's view) and half was totally functional. The "freaked out" went away, but the intensity and pace was always there. My boss

was Megan Smith, and her boss was the President. We were policy advisors, not operational. The CIO is responsible for Government IT cross-working, ... Three missions

1. improve the tech capbiity of the Government
2. improve the nation's capability to build and use technology (education, R&D etc.)
3. ensure public policy decisions are informed by the best technology advice. Note *all* policy decisions. For example, DoT Secretary, never met technology in confirmation etc., but were everywhere. Pittsburgh has lots of Uber cars. Self-driving trucks are being tested. In aviation, a major issue is drones. Verycomplex regularoty and safety issues. When Fox came into office, there was no regulatory infrastructure. There are multiple companies testing flying cars. One of the things I worked on, with Fox's team and others, was self-driving. Every year, kills 36,000 humans. People are terrible at driving. Its pratcically a certainty that machines will make much better drivers. The policy questions isn't "whether", it's "how do we get there", and "how do we get there soonest". Worked on Federal Automated Vehicles Policy. Needs a large team, e.g. software people, lawyers, regulatory experts. We *did* have a strong cohesive team, and this is one of the unheralded successes of Obama.

We are always on duty. Example San Bernardino attack. I worked on the technology part of the President's speech. I spent many hours on a few words. Then a long follow-up with technology leaders. Humorous story about how this "secret" meeting was blown by an "unnamed senior official" making a Starbucks run, captured by, *inter alia*, the Guardian's photographer.

Deputy SecDef on Third Offset strategy. First Offset was nuclear deterrent (1950s). Second was Stealth etc. from 1970s. We now need a Third, driven by technology: AI, ML and technology. Shows a US military drone, *not* autonomous. But clearly we could go much further. The logic leads to the vehicle using its computers to slect targets. Complex legal/ethical issues here. DoD policy on autonomy on weapon systems. Insists on compliance with international humanitarian law. When can a machine make the decision that someone is a combatant? What if a machine is better than a human at making this decision? How can we tell? What about our opponents? What about non-state actors, who might not adhere to international law. Also, much concern about ??.

Note that I actually worked in the Eisenhower Building. Built in 1870/80s, as the largest office building in the world, in French Empire style. Mark Twain: "ugliest buidling in America". But I disagree, and many now agree with me. When IBM etc. says "it's reinventing itself round AI", Government listens. We kept worrying about "The Robot Apocalypse". Held five conferences round the country in seven weeks. Produced a document "Preparing for the Future of Artificial Intelligence", reviewed by every Department, and many agencies and

councils, all of whom were helpful. Separate report on "How AI will affect the economy and job market". 3M Americans (2% of work force) work as drivers, and 2M are at risk through "self-driving". Graph on labour force participation. Women's increate 1950–2000. Men decline from 87% to below 70%, and indeed for women, it's been declining from the 2000 plateau. Aging is not enough of a driver. Stagnation in median[1] wage for several decades. factors include automation, globalisation and tax policy. Automation is not just tightening bolts, it is radiographers etc. Howver, the data he showed were themselves produced by ML, displacing a number of economics graduate students in the process.

Let's not be too pessimistic. AI could drive productivity growth, and making a larger pie *could* with the right policies, make everyone richer. We can absorb large changes, but it causes real pain in the process. There will be a lot of decisions, with technology input and ramifications, to be taken. Therefore we should step up: CS students, Uncle Sam [and Rosie the Riveter] need you.

[1] JHD was pretty sure he said "median" and not "mean", which is a key distinction. See the graph in `https://en.wikipedia.org/wiki/Household_income_in_the_United_States#Median_inflation-adjusted_.28.22real.22.29_household_income`.

Chapter 5

14 February 2017: Building Websites with Jekyll and GitHub

An ACM Student Chapter talk at NYU. `https://foureyes.githib.io/acm-github-pages/`. Speaker a "Clinical Assistant Professor" at NYU, who does all his course material this way.

Polled the audience: a lot of Git, shell experience, not much Ruby/Gem. Almost no markdown and Jekyll.

We'll start the easiest way, then gradually use more. A dogfood site, but also uses bootstrap and grunt for his slides etc.

So you want a website? How 1990's: why not just create a Facebook page? Talks through various options: one is self-hosting, but the speaker strongly advised against this (security being a major issue). Serious alternatives might be Neocities, Weebly, GitLab pages (pretty similar). netlify, Firebase (known as a database, but will do static).

5.1 Git hosting

Typically `https://username.github.io/reponame` or `https://username.github.io`. You can configure a custom name, but without https.

In your repository, you might want a `.gitignore`. Common convention is to use a gh-pages branch.

Shows markdown `__` bold, `_i ... _` italics, a link `[nyu](http://www.nyu.edu)` Inline code ‘`...`‘

Need the following at the start of the file.

```
---
layout: default
---
```

GitHub has added some extra features to Markdown.

5.2 Jekyll

One problem with the web interface is that it takes time. We should switch to developing locally. This also lets us test locally. The magic is that pages get run through a static generator which converts the Markdown to HTML. This is Jekyll. So to test locally. If something goes wrong with the build, you get an e-mail to the GitHub-registered address.

He used `gem install jekyll` then `jekyll build` This gives anew folder `_site`.

You can create an `_layout` folder. This can be used to create standard navigation across all pages etc.

Q Have you used ??

A I'd like to, but haven't.

Q Can you use pre-built templates?

A You can, but I don't.

His real repository has a `_css` folder, which in fact he doesn't write, but has generated by `grunt`.

Chapter 6

Akamai

Spike caused a 517Gb/sec DDoS in the third quarter, this is the third wave of DDoS (after Anonymous at 80Gb/sec using used PCs, organised groups as 300Gb/sec using servers), using enslaved IoT devices, and we've seen botnets with millions of devices. Note that "proof of existence" encourages others, e.g. 4-minute mile. Note that all these sizes are more than most data centres can withstand. The Dyn Managed attack (October 2016) was 1200GB/sec.

Note that Mirai (Japanese name) has many features: 62 default username/password pairs, flawed spoofing but interesting command and control. We now define a Mega-attack to be > 300, and we have seen $8 \geq 15/12/2016$. Graph of NTP attacks. "The bad guys were competing for resources". NTP reflection attacks seem to be drying up: down 27% from 2016Q3 to 2016Q4. But SSDP reflectors went up 321%. SSDP is Bluetooth self-discovery — every time you do this you open up other options!

One botnet has 12936 members, targeting 123909 accounts. Each machine was only logging on once every two hours: below most people's radar. This sort of botnet can only be spotted by multi-site visibility. Therefore you need cloud security providers (e.g. us) to detect these.

We are now seeing fewer web application attacks: peaked in Q12016. 51% SQL injection, 37% Local File Injection, 7.16% XSS. Most attack come from US, but UK, NL, DE also big. Detailed graph of the Thanksgiving period (big shopping days). EMEA saw very little spike then (visible but not large). In October 2016 was a big spike, but that was one attack. Big spike in EMEA around Xmas 2016. Large spike on Department stores. Notes that distribution of sources and targets are not the same, and there's a nice graph of cross-country flows: most of RU attacks are on Italy. Careful note on attribution: this is where the attack packets are coming from, so all we are seeing is the location of the botnets/proxies, not necessarily the "mastermind" location.

Things are changing: FTC took D-Link to court over lack of security. Also showed the "hackers kill a Jeep' story.

Q Why are you advocating regulation?

A Only reporting.

Q Are Chinese devices more vulnerable?

A As the number of devices increases, so certainly the number of complaints go up. Not just security, lead in paint etc. As a percentage, there probably wasn't a rise. Note also that China is no longer top of the source of attacks, even though there are more IoT devices. [JHD wasn't sure about the logic here.]

Q Cooperation?

A Yes, forums etc., also *ad hoc* collaboration (based on plans) for specific attacks.

Q SQL injection?

A Explanation.

Chapter 7

What's new in Maple 2017

7.1 Preliminaries

Planning release in May 2017. β-testing is now starting. Recommends Help $\rightarrow$ what's new.

7.2 Cloud Package Management

Cloud has been around for a while. Replaces Maths Group. Can also have a private cloud.

Maple has had packages, and user-written packages. The difference in 2017 is we are now encouraging people to put packages in the cloud. Right-click and do "install package". There's a new package called `PackageTools` for supporting this. Still needs a verb+with+ command as usual. Notification of updates is on the to-do list. Aim to develop the webview side, and ultimately this will replace the "Cloud" bar. Note that this adds to the library path. packages go into "Maple/toolbox".

Also, lots on new content in "Maple portal for Engineers". Also Geonames dataset. Note that these are plots, and support `PlotAnnotation`. In particular, can annotate curves on a contour plot. Statistics adds `VennDiagram`, `ViolinPlot` etc.

Q What if a package conflicts with the Maple library?

A Library has priority. We haven't quite formalised our rules about what we'll allow into the public cloud.

Workbooks can allow password-protected content. Hence educators can add code to workbooks that students can't see, but does run. Code generation now supports Swift.

Better handling of `assume`, `CouldItBe`, and known-on in integration.

There'll be another one of these end-April, early-May.

Q How do you delete a package from the cloud?

A right-click Delete.

Q Can we add web documents to the cloud?

A Maple worksheets/books, yes.

Q What's changed in limits?

A Look at the page — I can't really describe.

Can always get to `PlotBuilder` via the menus. Not also have `PlotExplorer`. This is our new interaction model. Hopefully it will be documented by β_2.

Chapter 8

Organization and Analysis on Data Tensors

8.1 Abstract

This Thursday we are honored to have Ronald Coifman (Yale) in the Mad (Math and Data) seminar.

Abstract: Our goal is to illustrate and give an overview of various emerging methodologies to geometrize tensor data and build analytics on that foundation. Starting with conventional data bases given as matrices , where we organize simultaneously rows and columns , viewed as functions of each other . We extend the process to higher order tensors,on which we build joint geometries.We will describe various applications to the study of questionnaires , medical and genetic data , neuronal dynamics in various regimes. In particular we will discuss a useful integration of these analytic tools with deep nets and the features they reveal.

Speaker recalled past regional seminars, and suggested another in this area

8.2 Science of Data

What are the mathematical structures we can put on these? When we classify, we tend to build an empirical model of what the data tell us, and then have some measure of "success'. Image of a "mandrill" as orchestrated by François Meyer. Take the wavelets. Then the residual is interesting, which can be modeled by brush strokes. The residual of this can be modeled as pointillistic. See our paper [PNAS2017] (probably [YTCK16]).

Looked at "Coupled pendulum" Can we reconstruct from a scrambled camera? Then "empirical physics" Gives us $\cos\left(\sqrt{\frac{g}{L}}t\right)$. Also takens–Bogdanov dynamical system: suppose we didn't observe the actual solution, but only a function of it. Then our model still describes the different states. So how does

one do this?

Example 3 *Data where each column is a yes/no response to 570 questions in a bizarre (but common) questionnaire.* 3000×570*: how do we make sense of this? Might want to correlate people, or correlate questions.*

Complex oscillatory phenomena (acoustic scattering off an ellipse) and he shows the results. Could also be obtained by local SVD.

One of the first applications of wavelet bases was the observation that CZ operators be represented as ....

When we build this geometry of the data, into a tree of subsets, then we get the construction of Harr wavelets on intervals. Then the question is how well we represent the data in this tensor basis. We can iterate this geometry until we can no longer reduce the entropy of the tensor-Harr expansion of the data base.

A deep net is basically doing gradient descent. The geometry of the data is defined by the functions you use to approximate it. See [YTCK16].

Chapter 9

Randomized Algorithms Meets Formal Verification Justin Hsu, University of Pennsylvania

9.1 Speaker's Abstract

Algorithms and formal verification are two classical areas of computer science. The two fields apply rigorous mathematical proof to seemingly disparate ends—on the one hand, analyzing computational efficiency of algorithms; on the other, designing techniques to mechanically show that programs are correct.

In this talk, I will present a surprising confluence of ideas from these two areas. First, I will show how coupling proofs, used to analyze random walks and Markov chains, correspond to proofs in the program logic pRHL (probabilistic Relational Hoare Logic). This connection enables formal verification of novel probabilistic properties, and provides an structured understanding of proofs by coupling. Then, I will show how an approximate version of pRHL, called apRHL, points to a new, approximate version of couplings closely related to differential privacy. The corresponding proof technique—proof by approximate coupling—enables cleaner proofs of differential privacy, both for humans and for formal verification. Finally, I will share some directions towards a possible "Theory AB", blending ideas from both worlds.

9.2 Preliminaries

Randomised algorithms are everywhere: theory and practice.

- High probability guarantees

- average case behaviour
- learning theory
- game theory
- cryptography
- Privacy, notably differential privacy.

We wish to establish these rigorously. Sparse Vector Algorithm satisfies differential privacy (1 ACM page proof), but there are six variants, some of which have subtle bugs. Can we add formal verification? Proofs a re subtle, and quite condensed. Also use had tools from probability theory.

9.3 New techniques for formally verifying randomized algorithms

Coupling is a way to correlate two distributions. Only links events with equal probabilities.

Definition 1 *A coupling of $\mu_1, \mu_2 \in Distr(A)$ is a joint distribution $\mu \in Distr(A \times A)$ such that the μ_i are projections.*

For example, if coupled samples are equal, $\mu_1 = \mu_2$, and so on. Hence a *coupling proof* is a proof that works by constructing a coupling.

Example 4 *For T iterations, flip a coin and more up/down by 1 depending. Consider two walks starting at k and $-k$. Consider the case where if the positions are different,we couple opposites, if the walks are in the same place, couple equal steps (and hence from now on they are always equal). Hence if the walks haven't met, they're symmetric.*

pRHL can be used to combine two programs into one.

9.4 DP

Use apRHL — Approximate equivalent of the above.

Q For this post-condition to make sense, doesn't the program have to have run? So what's the insight?

A Hoare-like logics always assume a program [terminates].

How does your proof technique deal with convolutions.

Q How does your proof technique deal with convolutions.

A Suppose I had two Databases that differed in one element, the the convolution would tell us ...

This proof style is easier for a computer to check.

9.5 Also

Use composition in differential privacy. Use incentive properties for universal truthfulness. Also martingales. In private equilibrium computation, we use two-player games and private query release as a technique. Have published in

Handbook of Theoretical Computer Science: Two volumes. Theory A; quantitative properties, impossibility results. Theory B: .... Note the two have split, whereas Church, Turing worked in both.

9.6 Future Work

What is the "right" definition of "approximate coupling" – several incomparable definitions currently. How does one go from slick proofs to formal verification? Can this also produce easier proofs for humans to understand?

Q Better tools — what can you offer, given the major developments in the area, Coq etc.?

A The space of theorem provers is quite large. I am envisaging a specialised tool for random algorithms. It's very tedious to build such proofs in generic tools.

Q What is the space to which your techniques apply?

A Imperative programming languages, with a "random sample" command. It is not clear exactly what the boundary is.

Q What are the problems with the Constructive Lovász Lemma?

A (JHD didn't follow)

Q Loops?

A If the loops are synchronised and we have exact couplings, this is pretty standard Hoare logic. Asynchronous/exact can also be handled. Approximate is more difficult.

Q Temporal logic?

A I don't know enough to be sure

Chapter 10

Hardening Cloud and Datacenter Systems Against Configuration Errors

10.1 Abstract

Configuration errors are among the dominant causes of service-wide, catastrophic failures in today's cloud and datacenter systems. Despite the wide adoption of fault-tolerance and recovery techniques, these large-scale software systems still fail to effectively deal with configuration errors. In fact, even tolerance/recovery mechanisms are often misconfigured and thus crippled in reality.

In this talk, I will present our research efforts towards hardening cloud and datacenter systems against configuration errors. I will start with work that seeks for understanding the fundamental causes of misconfigurations. I will then focus on two of my approaches, PCheck and Spex, that enable software systems to anticipate and defend against configuration errors. PCheck generates configuration checking code to help systems detect configuration errors early, and Spex exposes bad system reactions to misconfigurations based on configuration constraints inferred from source code.

Bio: Tianyin Xu is a Ph.D. candidate in Computer Science and Engineering at UC San Diego. His research interests intersect systems, software engineering, and HCI towards the overarching goal of building reliable and secure systems. His dissertation work has impacted the configuration design and implementation of real-world commercial and open-source systems, and has received a Best Paper Award at OSDI 2016.

10.2 Background

So much relies on cloud and data services. look at 2012 Amazon failure taking down 70 online services. Claimed 50% of cloud services have 3+ outages on headline news every year. Google's data (2013) shows 25% configuration errors, software at 30%, whereas hardware and network are each 10%. So why can't systems tolerate these errors. These errors tend to be epidemic, as they are propagated to all software instances, and quite often the error is the configuration of fault-tolerance.

So how do we built a data center robust against these? Have to understand the real-world first. Input from industrial collaborators. I want to harden the software systems. Paper at CHI17 on regarding configuration as an interface. I have exposed 780 misconfiguration vulnerabilities, of which 386 are confirmed /fixed. I've helped Squid improve their configuration library.

10.3 Systems approaches

10.3.1 Configuration Hell

as practitioners call it. The problem is the complexity. Hadoop has 312 parameters, and grows over time (Good graph). There are also lots of dependencies. MySQL 461, Apache 487, Storage-A 412. Administrators have no way to understand all these.

Looked at 620 cases. 27.7% were difficulties understanding, 58.5% errors, and 13.7% were other. Of the errors, 22.1% were default-setting errors, and 35.4% (rest) were customised. Note that the code itself is not buggy! In general, code does report these errors, but possibly not early enough. Also one wants useful reporting and graceful degradation.

Example 5 `signal(SIGSEGV,call_techsupport)` *but* `call_techsupport` *uses the* `dial_program`*, which goes to a configuration file. What if this is wrong?*

10.3.2 PCheck

Phases of a service: initialisation, rollout, workload and error. The later the phase, the greater the cost. What are the RAS features in common software. e.g. Apache has 14. MySQL has 43. 5–39% of these are subject to latent configuration errors: latent time bombs. Tool called PCheck auto-generating checking code for configuration parameters. These checkers should be invoked at initialisation (and periodically after!). Note that the checking logic already exists, e.g. in `dial_program`. Currently only invoked when `dial_program` is invoked: too late. Hence PCheck has to extract the code from where it is, and give it an appropriate context. Also side-effect problems (e.g. actually dialling tech support). To extract the code, we use static taint checking. Find the context by backtracking the taint-propagation process. To prevent side-effects, PCheck uses sandboxing.

Q What if context is huge?

A This is a best-efforts basis!

Different programmers have different conventions for mapping configuration variables to program variables, but we can't really expect developers to annotate every variables. There are three patterns: structure/getter/comparison. Look at first: PostGres(?) has one, holding 82 configuration variables. Only need to be told where this is. In a dataset of 58 real-world latest configuration errors (21 historical, 37 new ones found manually by us), we generated code with PCheck, and found 71% of historical, 78% of new errors. We got 100% of invalid data cases, and most invalid file names. The "miscellaneous" category (especially resource exhaustion) were the hardest. The cost was less that 1000msec, often < 100. Out of 830 configuration files, only got 3 False Positives.

10.4 Graceful

What use is a stack backtrace to the average administrator? Showed one, and explains how Microsoft fixed the error message to administrator-friendly. Hence we (developers) need to anticipate configuration errors. Again, note that we can infer constraints from the way the parameters are used. Hence we need misconfiguration injection (into the developers' world). Our tool is Spex for doing this.

10.5 Next

Look at security misconfigurations. Recall 40,000 misconfigured MongoDB databases were openly available. Roots in the obstacles to understanding (CHI'17). Understanding is even harder here. Note that Facebook has moved from "move fast and break things" to "move fast on stable infrastructure".

Q What is the relationship with debugging?

A Long answer, which JHD didn't really follows.

10.6 Subsequent note

JHD observed [Pag17], which describes an outage at Amazon Web Services that took out many IoT devices, including at least one person who couldn't top *off* his oven. There's a technical description at [Ama17]. It looks like this error is not the sort that would have been caught by the speaker's proposal, but the general point about dependence is very well made.

Chapter 11

A New Approach to Network Functions

Speaker: Aurojit Panda, University of California Berkeley

11.1 Abstract

Modern networks do far more than just deliver packets, and provide network functions – including firewalls, caches, and WAN optimizers — that are crucial for scaling networks, ensuring security and enabling new applications. Network functions were traditionally implemented using dedicated hardware middleboxes, but in recent years they are increasingly being deployed as VMs on commodity servers. While many herald this move towards network function virtualization (NFV) as a great step forward, I argue that accepted virtualization techniques are ill-suited to network functions. In this talk I describe NetBricks — a new approach to building and running virtualized network functions that speeds development and increases performance. I end the talk by discussing the implications of being able to easily create and insert new network functions.

Bio: Aurojit Panda is a PhD candidate in Computer Science at the University of California Berkeley, where he is advised by Scott Shenker . His work spans programming languages, networking and systems, and his recent work has investigated network verification, consensus algorithms in software defined networks and frameworks for building network functions.

11.2 Talk

Intersection of Networking, Systems and Programming Languages.

11.2.1 Evolution of networks

1. Standard diagram: hosts that do everything except packet delivery: switches and routers that do that.
2. Rise of middleboxes: security, caching, SSL proxies etc. Many manufacturers, and a survey showed that about 1/3 of all devices were middleboxes. These were originally ASICs, but have become general-purpose hardware. Faster hardware enabled this.
3. Note also the rise of virtualisation and virtual switches.

* So virtualise the middleboxes

4. Network Function Virtualisation. AT&T will virtualise 75% by 2020. This gives reduced cost and greater agility for carriers.

* So networks can evolve at the speed of software, rather than hardware (and persuading vendors to make boxes. But at the moment the concepts are still dictated by what the old vendors supplied.

Deployment is hard, because we want both isolation and performance (at line traffic rate: 10–100M packets/second) going, possibly, through several middleboxes. Traditional middlebox vendors have hand-optimised the performance. Hence many people can only deploy one NF middlebox per physical server, which rather thwarts the aim of virtualisation.

11.2.2 NetBricks

[OSDI'16] Three contradictory goals.

- High performance.
- Consolidation (many NFs on one physical machine).
- Isolation. Various meanings:
  - Memory isolation (NFs might have secrets, such as SSL keys);
  - Packet isolation (only one NF at a time has access);
  - Performance isolation (not really treated today).

For memory isolation, we use memory management, and copy packets (performance hit!). Could get 25Mp/s, but using OVS VM, this dropped to 4Mp/s. Our BESS VM gets this to 8Mp/s. BESS containers moves this to 10Mp/s. A 0-copy container is 15Mp/s.

If you run on different cores, you have to have a cache-cache copy, if not, you have context switches on the core. Both cost!

NetBricks relies on compile-time isolation. Note that we can't rely on garbage collection. We also can't do memory segregation since tracking processes costs too much. Note that packets live in raw DMA buffers typically. Rust provides type checks, bound checks without GC.

Also need packet isolation. Typically done by copying, which is the problem. Solution: *linear types*. The syntax marks arguments that are moved. Linear types are implemented by Rust for concurrency. NetBricks operators consume packet references. API is designed so that safe code can never learn packet buffer addresses.

All NFs run in one process space, in a linear chain of packet handling. below this there's DPDK polling for I/O, scheduler and NICs. The NF chain is not pre-emptable. This reduces the number of packets in flight, so working set. This allows NFs to share a core without context switching: only a function call (or less if we allow inlining).

Q Why is context switch so expensive? TLB?

A Some enforcement of isolation is necessary.

Q Dynamically add network functions?

A Currently, need to recompile. Future?

Three scenarios

- VM
- NetBricks
- NetBricks multicore (#core=#NFs)

With 1 NF, we see $\times 1.5$, 8NF is $\times 9.5$. All this on 60-byte packets, but benefits still there for larger packets.

Q Proof-Carrying Code?

A Type checking is a special case of PCC.

Q What can't you do?

A Lots of NFs use the new (Intel) AES-NI instructions, which Rust currently doesn't support. Tasks like flow identification are difficult. In general, various micro-optimisations, and global optimisations like cache coherence.

Q Industry?

A I went to "them" (?) to talk, but can't find out how they do it?

11.3 Opener Carrier Interface

Third parties can specify the processing graph. Automatically installs (enough copies of) NFs at the edge. Prototype implementation implementation.

11.4 Novel use case: verifying Microservices

The application developer writes down assumptions about the MS. Then it is verified that the MS satisfy these, and independently, the application is verified against these models.

Chapter 12

Domain Decomposition: From Poisson to Coupled Multi-physics Problems: Xiao-Chuan Cai

The Olof Widlund retirement lecture. "I was Olof's 13th student, one of the first doing domain decomposition (85–89). At that time, both parallel computing (Jack Schwartz) and domain decomposition were started. Used Cyber 205 (double/single/half precision)[1] at Princeton. Two important (unpublished!) papers: Schwartz and Widlund — showed Courant Tech Reports." To overlap or not to overlap? Paper in 1988 saying it doesn't matter much, but the debate is still going on. OW has 90 DD papers, 21 DD research students. Both classes of methods have developed, and DD and MG theories have, more or less, converged. PETSc is a large piece of software. OW's book is in use across the world.

12.1 Current State of Art: Non-overlapping DD

[KLR16] Goes from 100% efficiency on 8Ki cores to 96.3% on 768Ki cores.

12.2 Current State of Art: overlapping DD

[Yangetal2016, Gordon Bell Prize winner] Newton–Krylov–Schwarz scales out to 10M cores.

[1] Commented how useful the range of precisions was.

12.3 Current state of large scale PDE solvers

Elliptic: yes/no; mesh structured/non. A structured mesh means that we can define a coarse mesh.

E/S MG — essentially solved.

E/U AMG

NonE/S many level DD. See two previous sections.

NonE/U ?? (at least in practice).

Example 6 (Computational Biomechanics) *Blood flow in a human artery. The calculations is multiphysics, and requires unstructured meshes. We want to model the flow, and the wall shear stress (which is what causes the artery to break, and hence a stroke). These can't (easily) be captured by imaging. Note also that imaging can't do "what if", e.g. "what if the patient is active, rather than lying on the MRI machine?".*

Shows image of his own cranial artery. I extracted the geometry. They also measured flow rate at the neck, hence the boundary conditions. I also know viscosity and blood pressure, so that's the rest. Can compute "efficiency" and it comes out at 86% (95% for healthy youngsters). 70% is the (empirical) critical number. Shows video of a simulation of the blood vessels in his head.

Similar techniques for heart and lungs. "Generally speaking, CFD is not yet popular in medicine, unlike aerospace or automobile." HeartFlow at Stanford, CHeart at KCL are used clinically. Doing joint work with Stroke Center, Beijing Tiantan Hospital (hard to find good collaborators!). Manged to find seven patients to do calculation versus measure. They managed to measure the pressure at two points. Has a graph: worst discrepancy 20%, but once exact. Our calculation is at the midpoint of the artery, the sensor might drift. The doctors now believe that the calculations are valid.

12.4 Fluid-structure interaction problems

Two families.

Iterative Solve the fluid structure equations, update each other's boundary conditions and iterate. This is what you get in the ANSYS's of this world. Convergence is very difficult.

monolithic (fully coupled). Solve fluid, structure and mesh movement equations simultaneously

Let u_f and p_f denote velocity and pressure of the fluid. Use N-S with an additional term for moving wall. In small arteries need a non-Newtonian version. Elasticity equation for the wall, with a damping term to prevent the artery from

becoming arbitrary large. Saint Venant–Kirchoff method. The fluid domain displacement at time t assumed to satisfy a linear elasticity equation. Also need boundary conditions at the neck.

Implicit FE discretization. $\mathcal{F}$ is highly nonlinear. Convective term is a major cause. Basic preconditioner is a One-level restricted additive Schwarz preconditioner. Subdomains obtained by graph partition, and each subdomain might have flow variables, or solid variables, or both, Works for a few hundred processors, then we need more levels. 19-line algorithm for multilevel multiplicative Schwarz (V-cycle). Actually additive within levels and multiplicative across levels. Hard problems also need smoothing. For me, DD doesn't care about smoothing, but MG has to worry.

For elliptic problem, Xuejen Zhang has the theory in 1991. Most people use two levels. or structured meshes, generating the multiple coarse meshed (3 parameters per level), and can stage to 10^6 processors For unstructured, generating the coarse meshes is far more difficult. Scaling beyond 10^4 is hard. In practice, communication dominates in these cases.

A coarse mesh must be very small, and that implies removing points from meshes that are already coarse. The algorithm first selects a subset of vertices S on the fine mesh to preserve the boundary geometry, Have to respect the geometry of the interface. This is easy to say, hard to write the code! "Isogeometric" is the key. Keep all the points on the interface, but remove interior points.

Pulmonary artery. There's a complex branching. 10Ki processor scaling. (63–93%) efficiency.

I need more data — will you be my Ω? Get a copy of your MRI, and I can do the rest.

Q–OW mentioned various alternative approaches.

A Yes, but issues of accuracy.

Q Does the pressure probe change the flow?

A Yes, we allow for that.

Q What about if there's a clot?

A We can perturb ....

Q Smoothing?

A Done at the pre-conditioning step only.

Chapter 13

Compositional Models for Information Extraction

13.1 Abstract

Monday February 27, 2017 2:00 P.M., 60Fifth Ave, Room 150 Compositional Models for Information Extraction Mark Dredze, Johns Hopkins University

Synopsis: Advances in machine learning have led to new neural models for learning effective representations directly from data. Yet for many tasks, years of research have created hand-engineered features that yield state of the art performance. This is the case in relation extraction, a task in the field of information extraction in which a system consumes natural language and produces a structured machine readable representation of relationships between entities. Relation extraction systems are the backbone of a many end-user applications, including question answering, web search and clinical text analysis.

I will present feature-rich compositional models that combine both hand-engineered features with learned text representations to achieve new state-of-the-art results for relation extraction. These models are widely applicable to problems within natural language processing and beyond. Additionally, I will survey how these models fit into my broader research program by highlighting work by my group on developing new machine learning methods for extracting public health information from clinical and social media text.

Bio: Mark Dredze is an Assistant Research Professor in Computer Science at Johns Hopkins University and a research scientist at the Human Language Technology Center of Excellence. He is also affiliated with the Center for Language and Speech Processing, the Center for Population Health Information Technology, and holds a secondary appointment in the Department of Health Sciences Informatics in the School of Medicine. He obtained his PhD from the University of Pennsylvania in 2009.

Prof. Dredze has wide-ranging research interests developing machine learning models for natural language processing (NLP) applications. Within machine

learning, he develops new methods for graphical models, deep neural networks, topic models and online learning, and has worked in a variety of learning settings, such as semi-supervised learning, transfer learning, domain adaptation and large-scale learning. Within NLP he focuses on information extraction but has considered a wide range of NLP tasks, including syntax, semantics, sentiment and spoke language processing.

Beyond his work in core areas of computer science, Prof. Dredze has pioneered new applications of these technologies in public health informatics, including work with social media data, biomedical articles and clinical texts. He has published widely in health journals including the Journal of the American Medical Association (JAMA), the American Journal of Preventative Medicine (AJPM), Vaccine, and the Journal of the American Medical Informatics Association (JAMIA). His work is regularly covered by major media outlets, including NPR, the New York Times and CNN.

13.2 Talk

Began by showing a word-cloud from his published papers. 'model' and 'data' were the biggest. Co-founded the ML group at JHU. ML is now a service course offered every semester: was 40, now 80 plus 85 on wait list. My domain expertise is in public health. Talk about four areas.

13.2.1 Modelling Users

Consider Public Health campaigns, typically aimed at changing behaviour. Example: "great American smoke-out". Hard to evaluate. Followed Quitline calls, Twitter etc. There are also spontaneous campaigns, e.g. Charlie Sheen announces he's HIV-positive. This generated more interest than any organised HIV campaign.

Also model world-wide travel for epidemic forecasting. Zika virus is the test case. So we care about demographics. Have a Gated Recurrent Unit model for demographics from user names. Also use Generalized Canonical Correlation Analysis.

13.2.2 Multi-Factor Topic Models

SPRITE = Structured Priors for Topic Models. [PD15].

13.2.3 Compositional Representations

Q Why?

A Clinical data is used for a lot, but structured data. Most people don't use just text.

Q Turning text into actionable data?

A We do better by modifying the problem. Financial analysts don't need a general database. Limiting the data's context helps a great deal.

Q End-to-end solutions, e.g. a NN.

A In a more restricted domain, we can do better. Challenging to find the right balance, though.

Chapter 14

Gitta Kutyniok: Optimal Approximation with Sparse Deep Neural Networks

14.1 Abstract

Deep neural networks show impressive results in a variety of real-world applications. One central task of them is to approximate a function, which for instance encodes a classification problem. In this talk, we will be concerned with the question, how well a function can be approximated by a deep neural network with sparse connectivity, i.e., with a minimal number of edges. Using methods from approximation theory and applied harmonic analysis, we will first prove a fundamental lower bound on the sparsity of a neural network if certain approximation properties are required. By explicitly constructing neural networks based on certain representation systems, so-called α-shearlets, we will then demonstrate that this lower bound can in fact be attained. Finally, given a fixed network topology with sparse connectivity, we present numerical experiments, which show that already the standard backpropagation algorithm generates a deep neural network obeying those optimal approximation rates. Interestingly, our experiments also show that restricting to subnetworks, the learning procedure even yields α-shearlet-like functions. This is joint work with H. Bölcskei (ETH Zurich), P. Grohs (Uni Vienna), and P. Petersen (TU Berlin).

14.2 talk

Last years have produced great results from Deep NNs, e.g. Go, ImageNet, Siri. "Sparse" = "Sparse Connectivity", very few weights $\neq 0$.

d dimension of input layer

L # layers

N # neurons

M # edges

$\rho : \mathbf{R} \to \mathbf{R}$ Nonlinear function knows as *rectifier*

Theorem 1 (Universal Approximation) *Every continuous function can be approximated to $\epsilon > 0$ with a network with one hidden layer and $O(N)$ neurons.*

How are ϵ and N related? In approximation theory, the largest γ such that error is $O(M^{-\gamma})$ is the optimal approximation rate. Do the same were w.r.t. M. There's been a lot of work with one hidden layer, also with many.

14.2.1 Fundamental Lower Bound

Rate distortion theory. Let $E^l := \{E : L^2(\mathbf{R}^d) \to \{0,1\}^*\}$ denote the binary encoders of length l, ditto decoders D^l.

$$L(\epsilon, C) := \min\{l \in \mathbf{N} \exists E, D \dots$$

Theorem 2 (us) *Let Learn: $(0,1) \times C \to NN + \infty, \infty, d, \rho$ with ... (weights need at most $-c \log_2 \epsilon$ bits) $\epsilon^\gamma \sup_{f \in C} M(Learn(\epsilon, f)) \to \infty$.*

14.2.2 Realisation

Choose a set of function $C \subseteq L^2(\mathbf{R}^d)$.

Theorem 3 (us) *Assume .... Then there is a neural network $\Phi \in MM_{\dots}$ s.t. ....*

14.2.3 Applied Harmonic Analysis

JHD didn't follow this part.

Q What about high-dimensional objects, images? We find the sharing of weights matters: how many different ones.

A Good point. These results are still being written.

Chapter 15

Making the fastest routers programmable

15.1 Abstract

Friday March 03, 2017 11:30 A.M., 60 Fifth Ave, Room 150 Anirudh Sivaraman, MIT

Synopsis: Historically, the evolution of network routers was driven primarily by performance. Recently, owing to the need for better control over network operations and the constant demand for new features, programmability of routers has become as important as performance. However, today's fastest routers, which run at line rate, use fixed-function hardware, which cannot be modified after deployment. I will describe two router primitives we have developed to build programmable routers at line rate. The first is a programmable packet scheduler. The second is a way to execute stateful packet-processing algorithms to manage network resources. Together, these primitives allow us to program several packet-processing functions at line rate, such as in-network congestion control, active queue management, data-plane load balancing, network measurement, and packet scheduling.

This talk is based on joint work with collaborators at MIT, Barefoot Networks, Cisco Systems, Microsoft Research, Stanford University, and the University of Washington.

Bio:

Anirudh Sivaraman is a Ph.D. student at MIT, advised by Hari Balakrishnan and Mohammad Alizadeh. His recent research work has focused on hardware and software for programmable high-speed routers. He has also been actively involved in the design and evolution of the P4 language for programmable network devices. His past research includes work on congestion control, network emulation, improving Web performance, and network measurement. He received the MIT EECS department's Frederick C. Hennie III Teaching Award in 2012 and shared the Internet Research Task Force's Applied Networking Research Prize

in 2014.

15.2 Talk

Traditionally, we have programmable clients and servers, and fixed-functionality routers inbetween. This has been one of the strengths of the Internet's evolution. But these days there is no agreement on what should be in a router. Shows a timeline of functionality in routers: not many 1990s are still there. Measure aggregate capacity. Until 1985, the fastest routers were minicomputers. Since then, there's been a $\times$10–100 gap between software and hardware routers. Also, performance of software routers depends on what functionality is enabled.

Rather that baking in functionality, can we bake in primitives. Diagram of a router chip. Ingress pipeline $\rightarrow$scheduling queue $\rightarrow$egress pipelines. Classic ingress pipeline is Forwarding $\rightarrow$ACL $\rightarrow$Tunnels. Egress might be Measurement $\rightarrow$Multicast.

My goal is to build a fast router, not a general-purpose machine.

15.2.1 Programmable Pipelines: Domino

[SCB+16]: hardware primitives + compiler. Hardware pipeline is basically a set of matching patterns/actions. Very deterministic: one packet per clock cycle. Memory is local to action units, therefore access is deterministic.

We therefore build out of atoms: action unit + local memory. 1 packet/cycle (1 ns). The atoms

Compiler has a DSL: note that there can't be loops. Example of "sample every 10th packet". Typical algorithms are HULL etc. Showed Verilog for his atoms. Most atoms $< 1000\mu m^2$ at 32nm process. Nested conditional accumulate (for HULL, AVQ) was 3597. No floating point, so some algorithms requiring square root (CoDel [NJ12]) are not implementable.

Some hardware routers can do queue-peeking from ingress, which we don't support.

15.2.2 Programmable Scheduler: PIFO

[SSA+16]: why is the scheduler important? Depends what we're optimising for. A Cloud might want fairness between tenants (nodes), but a private owner might want minimising something else. The current state is a set of hardwired algorithms. There's no consensus on the primitives. Decides order of packets, and what time (flow limits) it is sent. First idea is a programmable dequeue function. But there's a very tight time budget (5 cycles at 100G). Therefore we need to refactor the scheduler. In many schedulers, the relative order of buffered packets does not need to change. Hence a "push in first out queue", which is essentially a priority queue. Hence what's programmable is the computation of the packet's rank. We run the rank computation in the ingress pipeline.

Example 7 (Fair queueing) *Rank is virtual start time.*

Example 8 (Token bucket shaping) *If there are enough tokens, send now, and rank = time now + amount of token waiting.*

Hierarchical Packet Fair Queueing violates this rule. But a hierarchy of PIFOs will solve this.

Our performance targets were typical for a shared memory router: 1GHz pipeline, 1k flows/physical queue, 60K packets (12MB buffer, 200 byte cell). the naive solution is a sorted array of 60K elements, infeasible. Ranks increase within a flow, so sort 1K of head packets: fine.

Other algorithms we can't currently handle are those that require f.p., periodic timer-driven computations, and egress-dependent ones.

15.3 Conclusion

Industry interest in PIFOs in FPGAs, Domino's packet transactions are now in P4 (an emerging router language).

Would like to move beyond routers to NICs, as faster networking requires more to be done here. Also interested in middleboxes. Even more generally, the classic "end of Moore's Law" means we need hardware specialisation.

Q What about "if it's too late drop it"?

A That can be done before the PIFO.

Chapter 16

Towards Automated Machine Learning

Alp Kucukelbir, Columbia University; 20 March 2017.

16.1 Abstract

Probabilistic modelling is changing the way we do science. We want to study large datasets to shed light onto natural processes. (How do proteins function? How do social networks form?) To this end, we need tools to rapidly and iteratively explore hidden patterns in data. However, using probabilistic models to infer such patterns requires enormous effort and cross-disciplinary expertise. My goal is to develop easy-to-use machine learning tools that empower scientists to gain new insights from data. In this talk, I will describe some of my recent research in new mathematical approaches to automating inference and building effective probabilistic models.

Alp is a postdoctoral research scientist at the Data Science Institute and the department of Computer Science at Columbia University. He works with David Blei on developing scalable and robust machine learning tools. He collaborates with Andrew Gelman on the Stan probabilistic programming system. Alp received his Ph.D. from Yale University, where he was awarded the Becton prize for best thesis in engineering and applied science. He holds a B.A.Sc. from the University of Toronto.

16.2 Talk

Machine learning is powerful, but hard to use. Define ML as a process that produces hidden patterns in data, and expose causal[1] relationships.

[1] His words; JHD would disagree.

Example 9 *A protein. 5.5M voxels. Looking at the places where the protein folds.*

Example 10 (LDA on arXiv abstracts) *1.9M words.*

Example 11 (Taxi rides in Oporto) *A public database with GPS data.*

Use statistical models, capturing uncertainty via probability. Takes an ML expert (PhD student!) months to actually produce results. I want to replace the PhD student with an algorithm ADVI. If we produce answers fast enough, we can refine the model and iterate.

> Statistical models are built iteratively as we build a model, ... [Box1960]

Consider Bayesian models Likelihood $p(x|\theta)$. Prior is $p(\theta)$. θ are latent variables. Posterior $p(\theta|x) = \frac{p(X,theta)}{\int p(X,\theta)d\theta}$. This denominator integral is the bottleneck.

MAP Delta function at $\arg\max_\theta p(X, \theta)$

MCMC Gibbs, Metropolis–Hastings, ...

Variational

All algorithms are tied to the model $p(X, \theta)$.

Example 12 (Stan) *pPCA model. Groups of people take taxi rides in similar shapes along segments of highway. So write down a supervised pPCA. This would have taken months by hand. Had two clusters, essentially tourists and locals.*

$$\phi^* = \arg\min_\phi KL(q(\theta, \phi||p(\theta, X)).$$

but we can't solve this, so

$$\phi^* = \arg\max_\phi E_{q(\theta)}[KLp(X, \theta 9] - E_{q(\theta)}[\log(q(\theta; \phi)].$$

Automatic Differentiation Variational Inference. This supports any model with $\nabla_\theta p(X, \theta)$. which is a wide range (lists them). But, of course, not discrete models. Conditional conjugacy doesn't matter.

If my problem is on $\theta \in \mathbf{R}^+$, we use exp to map to $\zeta \in \mathbf{R}$, but need a Jacobian normaliser. Similarly for upper bounds, lower bounds, both, and for a simplex we use a stick-breaking model. Stan has a compiler to manage all this. In ζ-space, we use gradients by automatic differentiations, and then stochastic gradient descent.

16.3 Algorithm

Input Data X, model $p(X, \theta)$

Output Model.

16.4 Q&A

Q How many samples?

A We use a single MC sample per iteration, as we have a lot of information in the gradient. Can get trapped in local optima: working on diagnosing this.

Chapter 17

Umar Syed: Statistical Cost Sharing

An NYU Machine Learning seminar: see [BSV17].

17.1 Introduction

Example 13 (Motivating) *Attributing battery usage to apps on a smart 'phone. We don't have the detailed information, and also it's not additive, especially in two apps want GPS, only one copy is instantiated. Also, one app turns the radio on, and another app goes into overdrive as a result. I work on Android, and different engineers will give different answers.*

Given $f : 2^N \to R$ where N is a set of n applications. $f(S)$ is the battery usage for the set S of applications running. Goal is to find "fair" cost allocations to the elements of N.

Definition 2 *The literature assumes "value query', i.e. given S, one can query $f(S)$. But in practice we have samples (for a given device: different devices are different problems), i.e. pairs$S_i, f(S_i)$).*

17.2 Core

Definition 3 *[Gil59] $c_1, \ldots, c_n$ is in the core of f if for all $S \subseteq N$ $\sum_S c_i \leq f(S)$, i.e. S would be charged more if they went alone.*

[Balcanetal2015a] took a $poly(n, 1/\delta)$ set of samples and ....

Our approach:

1. compute costs that satisfy core property on samples

2. argue by generalisation bounds that they satisfy core property on new samples

3. Theorem: sample complexity to find vector in probably stable core $O(N/\delta^2)$ [Balcanetal2016a]

Note that we can talk about approximately stable core, i.e. $(1-\epsilon)\sum_S c_i \leq f(S)$.

17.3 Shapley values

1. [Efficiency] $\sum_N c_j = f(N)$
2. [Symmetry] If $f(S \cup \{j_1\} = f(S \cup \{j_2\}$ for all S disjoint from j_1, j_2 then $c_{j_1} = c_{j_2}$.
3. ...
4. ...

These are essentially unique.

Definition 4 (submodular) *f is submodular if the marginal cost is submodular: $f(S \cup \{i\}) - f(S) \leq f(T \cup \{i\}) - f(T)$ if $S \supset T$.*

So not the "radio example"

Theorem 4 *Submodular function with bounded curvature κ can be computed efficiently ... $\sqrt{1-\kappa}$-approximable.*

For additive functions, Shapley values are marginal contributions, and bounded curvature is close to additive.

17.4 Data-driven cost sharing

New axioms.

1. If j_1, j_2 never co-occur in my samples, $c_{j_1} = c_{j_2}$.
2. If j never occurs, its cost is 0.
3. [Additivity] $c_j^{\alpha D_1 + \beta D_2} = \alpha c_j^{D_1} + \beta c_j^{D_2}$.
4. [Efficiency] $\sum c_j = E[f(S)]$.

JHD objected that, if j never occurs, $c_j = 0$, but j_2 never occurs with jm so $c_{j_2} = c_j = 0$ by axiom 1. Speaker hesitated.

Theorem 5 $c_j^D := sum_{S:j \in S} P(S \sim D) \cdot \frac{f(S)}{|S|}$.

Compare with the plus/minus statistic in team sports.

Q Android battery usage tool.

A Doesn't use this yet!

Q What's a "good" definition?

A People think of a set of axioms, argue intuitively, then prove uniqueness. Many options!

Chapter 18

Addressing Computational and Statistical Gaps with Deep Neural Metworks: Joan Bruna

Courant Machine Learning Seminar.

18.1 Introduction

$$\theta^* = \arg\min_\theta F(\theta).$$

MLE, Model Selection, Supervised Learning etc.

But finite sample size, finite computational budget, finite signal/noise ratio. [BottouBosquet]. There are iterative methods.

$$\theta^{(n)} = f(Fm\nabla F, \theta^{(n-1)}, \theta^{(n-2)}, \ldots)$$

Nesterov, Statistic Average Gradient etc. When F is convex, we can ask for consistency $\theta^* = \lim \theta^{(n)}$, and always for convergence speed.

18.2 Non-asymptotic question

Given budget B, minimise $F(\hat{\theta}_B) - F(\theta^*)$

Theorem 6 (Fundamental Theorem of ML) [1]

$$
\begin{aligned}
E(\theta^*) - \min_G E\{(l(G(X), Y)\} \quad = \quad & E(\theta^*) - \underbrace{\hat{E}(\theta^*) + \min_\theta(\hat{E}(\theta) - \min_\theta E(\theta))}_{\textit{statistical error}} \\
& \underbrace{+\hat{E}(\theta^*) - \min_\theta \hat{E}(\theta)}_{\textit{optimisation error}} \\
& \underbrace{+ \min_\theta E(\theta) - \min_G E\{l(G(X), Y)\}}_{\textit{approximation error}} .
\end{aligned}
$$

Lassoo: convex, unique solution for generic D, not strongly convex in general. Gram matrix is degenerate.

Iterative splitting via a surrogate function. Sublinear convergence due to lack of strong convexity. [BeckTeboulle2009]. Looks like a neural network with V and ρ as layers. In theory infinitely many, but we want convergence. What about a shallower network with trained parameters? [Oymaketal2015]. There's a phase transition between #measurements and convergence rate of optimization. Also [Giryesetal2016] describes the tradeoff between accuracy and convergence speed.

Why does this work? Principle of proximal splitting. The regularisation term $||z||_1$ is separable the canonical basis. Consider a unitary matrix A.

$$E(z) \leq E_A(z; z^{(n)}) = E(z^{(n)}) + \langle B * z^{(n)} - y), z - z^{n)} \rangle + Q(Az, AZ^{(n)}).$$

[1] JHD is less convinced than usual by his accuracy here.

Chapter 19

Safeguarding Users from Adversarial Networks

19.1 Abstract

Wednesday March 29, 2017 2:00 P.M., 60Fifth Ave, Room 150 Roya Ensafi, Princeton University

ISPs and governments are increasingly interfering with users' online activities, through behaviors that range from censorship and surveillance to content injection, traffic throttling, and violations of net neutrality. My research aims to safeguard users from network interference by building tools to measure, understand, and defend against it. In this talk I will present Spooky Scan, a measurement technique based on TCP/IP side channels that remotely detects specific types of interference almost anywhere on the Internet. In contrast to previous approaches — which rely on volunteers in censored regions to deploy custom hardware or software — Spooky Scan achieves significantly better coverage, lower costs, and reduced risk to volunteers. I am working to deploy Spooky Scan and related techniques in Censored Planet, a system for continuously monitoring global Internet censorship.

I will also describe two studies on the Great Firewall of China (GFW). The first study explores how the GFW finds hidden circumvention tools; the second discovered a new packet injection attack carried out by the GFW. These studies can ultimately inform public policy discussions and improve censorship circumvention tools. By uncovering network interference, we can hold ISPs, governments, and other network intermediaries accountable, and develop better technical approaches for keeping users safe.

Bio: Roya Ensafi is a postdoctoral fellow at Princeton University. Her research focuses on security and privacy, with an emphasis on designing techniques and systems to protect users from hostile networks. She won the 2016 Applied Networking Research Prize from the Internet Research Task Force (IRTF) for her research on the Great Firewall of China. While earning her Ph.D. at

the University of New Mexico, she received the Ph.D. Dissertation Distinction Award, Best Graduate Student Mentor Award, and Sigma Xi Research Excellence Award. She is a native of Birjand, Iran and enjoys climbing, biking, and basketball.

19.2 Talk

Aim to detect network interference, understand the behaviour of interferers, then defend. Naïve internet model, then add firewalls, packet inspectors etc. Look at CIA[1] principles. Turkey shutdown the Internet in 2016, Iran throttling in 2013, Google has 703 instances [?in 2016] of Government requests for take-down. There's monitoring, targeting and even modification. It's a complicated landscape, different actors have different agenda.

19.2.1 Detection

Problem 1 *How can we measure censorship, more precisely, can two clients talk to each other?*

RIPE Atlas etc. deployed hardware or software at hosts. Not scalable. Also, this activity can be classed as espionage. World map of IPv4: 140M sites that respond. How can I leverage this? "Spooky Scan" is a TCP/IP application, relying on the following TCP/IP features.

- Recall three-way handshake.
- Recall IP ID field in headers: need that clients have a global value for this.
- Note also that an unexpected SYN+ACK triggers a RST, and a SYN will elicit multiple SYN+ACK until acknowledged/timed out.
- Need to be able to send spoofed packets from measuring machine.(Need to clear this with local gateways, typically)

Then Spooky Scan operates as follows.

1. Send SYN+ACK to server, get RST and IP ID.
2. Send spoofed packed to client, SYN pretending to be from server.
3. Send SYN+ACK to server, get RST and IP ID, which should have been incremented by phase 2.

In a noiseless model, I get deltas of 2,1,4.[2]

[1] Confidentiality, Integrity, Availability!

[2] JHD doesn't believe 4, which would imply three SYN+ACK acknowledging a SYN. It depends when step 3 happens, but if we waited long enough we'd get more than three SYN+ACK.

Q SYN cookies?

A I've never seen this get in the way?

But what about noise? One solution is to magnify the signal, i.e. do step 2 five times to get 6/1/16. Also can repeat the experiment to get a feel for change in IDs.

Example 14 (TOR relay in Sweden) *China blocked server-to-client, US not, Azerbaijan blocked both.*

Want a better way, observing that different servers have different noise.

1. For 30 second, query IP ID every second.
2. over 30 seconds, send 5 SYN/sec and query IP ID every second.

* Now replace above by sequential hypothesis testing.

We could find 22M machines with port 80 open.

Issue But mustn't compromise unwitting clients.

A Put the vantage point behind a couple of routers, rather than at the front. Even after this I have more than 50K observers in 180 countries. Also have 100+ volunteer activists.

Q Framing people?

A There are much easier ways.

19.2.2 Understanding

Focus on China.

Problem 2 (Who did DDoS on GitHub, 2015 March) *JavaScript maliciously injected into Baidu's traffic as it left China. Possibly because GitHub hosts* `greatwall.org`. *We named the tool Great Cannon.*

Worked out this methodology for GC.

1. Packet goes into GFW
2. If not Baidu leave alone
3. if Baidu, with probability 1/50, inject JavaScript.

We could tell that the Great Cannon was co-located with GFW, as both were two hops in.

Suppose GFW were looking for packets from a US defence contractor. We could find these, going to, say, a Chinese bank, and inject a zero-day.

Suppose I create a TOR relay, and only tell one friend in China the address. There are 6000 TOR relays in all. The TOR handshake itself is not encrypted, so can guess with Deep Packet Inspection it's ToR, but not with enough certainty. Hence first send an active probe for a TOR handshake first, to see if the target really is TOR. There are several such active probes.

19.2.3 Understanding

[Unix Security 2016] TOR is being blocked by IP/Port combination. So I built a Sybil infrastructure, forwarding ports 30,000 to 30,600 to a TOR port. This found 1090 IP probers. Mean delay between TOR connection and active probe is 500ms. 22 hours later, got another set of probes, testing whether the machine was still a TOR relay (good housekeeping if you're a blocking firewall).

19.2.4 Control over Routing

Joint work with Rexford and others. How can I say "my packet mustn't go through X". Brazil put in a direct cable to Portugal to prevent traffic going to US. But many `.br` sites are actually hosted in USA: cheaper. Also, not all Brazil ISPs used this cable (probably incompetence).

Hence we need an overlay network to control routing.

19.2.5 Conclusions

These middleboxes are getting cheaper to buy. We therefore need to understand this modality. Future work: to deploy Censored Planet. Study interference in other networks, e.g. IPv6. Also work on "soft interference", e.g. throttling in Iran's 2013 election.

In one election, we started monitoring the network traffic six months before the election. I want to do this consistently. Also note that encryption means that governments need malware to target activists [Bahrain example].

Also want to understand the motivation behind people's usage of circumvention tools. What works in which country?

Chapter 20

Probabilistic Deep Learning and Black Box Variational Inference

20.1 Abstract

Friday March 31, 2017 11:30 A.M., 60 Fifth Ave, Room 150 Probabilistic Deep Learning and Black Box Variational Inference Rajesh Ranganath, Princeton University

Abstract: Scientists and scholars across many fields seek to answer questions in their respective disciplines using large data sets. One approach to answering such questions is to use probabilistic generative models. Generative models help scientists express domain knowledge, uncover hidden structure, and form predictions. In this talk, I present my work on making generative modeling more expressive and easier to use. First, I present a multi-layer probabilistic model called deep exponential families (DEFs). Deep exponential families uncover coarse-to-fine hidden structure. These models can be used as components of larger models to solve applied problems, such as in recommendation systems or medical diagnosis. Though expressive, DEFs come with an analytical challenge—scientists need to compute the hidden structure given observed data, i.e., posterior inference. Using classical methods for inference in DEFs is tedious and impractical for non-experts. Thus, in the second part of the talk, I will describe my work on black box variational inference (BBVI). BBVI is an optimization based algorithm to approximate the posterior. BBVI expands the reach of variational inference to new models, improves the fidelity of the approximation, and allows for new types of variational inference. We study BBVI in the context of DEFs to fit complex models of text and medical records. Black box variational methods make probabilistic generative models and Bayesian deep learning more accessible to the broader scientific community.

Bio: Rajesh Ranganath is a PhD candidate in the Computer Science Department at Princeton University. He works on easy-to-use, flexible machine learning methods with David Blei and on machine learning for medicine with collaborators at the Columbia University Medical Center. He obtained his BS and MS from Stanford University in computer science. Rajesh has won several awards and fellowships including the NDSEG graduate fellowship and the Porter Ogden Jacobus Fellowship, the highest honor for doctoral students at Princeton University.

20.2 Talk

Discipline Knowledge $\rightarrow$ model $\rightarrow$ add data and fit the model $\rightarrow$ .... Might be calibrated model, might be predictions,might be express prior knowledge. Want to get answers quickly (fast model development, fast computation). One technique is probabilistic generative model.

- driven by discipline knowledge
- built from reusable blocks
- Focus on discovering structure in unstructured data
- Prediction.

My building block is *deep exponential families*. Create families of probability distributions that reflect the intuitions behind neural networks.

Example 15 (Motivating) *Healthcare and the Electronic Health Record. Both the personal data and the hospital data are complicated. Interest in survival analysis (time to pass to next state).*

Normal, multinomial, Poisson are all exponential. So if we stack any layers, what do we get. Graphical representation: shaded node is data. Open circles are unobserved variables.

$$p(x|\eta) = \ldots T(x) \ldots .$$

$T(x)$ is a sufficient statistic for the distribution. Two sources of nonlinearity: explicit, and the derivative of the normalisation function. Each latent variable can be viewed as a generalised linear model.

What we want a sparse overlapping components. Example from 300K patients. Spotted a sleeping medicine coming up in pregnancy cases: medically unknown. Gamma distribution

$$p(z) = z^{-1} \exp(\alpha \log(z) + \ldots)$$

Problem is estimating the posterior. [McCullaghNelder1999].

Example 16 *Word count of word i in document n is Poisson. NYT and* Science*. LDA and* Γ*, Poisson, Sigmoid, with 1,2,3 layers. Best is 3-layer* Γ*.*

Q How does this differ from root parameterisation.

A RP works only for differentiable.

Q Model checking?

A This is an interesting question. In one sense all models are wrong. The question is how well the distribution matches new data.

Chapter 21

Revolutionizing Medicine through Machine Learning and Artificial Intelligence

21.1 Abstract

Mihaela van der Schaar, UCLA, on sabbatical at University of Oxford and Alan Turing Institute

Abstract: In this talk, I will describe some of my research on machine learning for personalized medicine. Because of the unique and complex characteristics of medical data and medical questions, many familiar machine-learning approaches are inadequate. My work therefore develops and applies novel machine learning methods to construct risk scores, early warning systems and clinical decision support systems for screening and diagnosis and for prognosis and treatment. This work achieves enormous improvements over current clinical practice and over existing state-of-the-art machine learning methods. By design, these systems are easily interpretable and so allow clinicians to extract from data the necessary knowledge and representations to derive data-driven medical epistemology and to permit easy adoption in hospitals and clinical practice. My team has collaborated with researchers and clinicians in oncology, emergency care, cardiology, transplantation, internal medicine, etc. You can find more information about our past research at: http://medianetlab.ee.ucla.edu/MedAdvance.

Bio: Mihaela van der Schaar is the Man Professor, Oxford-Man Institute, Department of Engineering Science, University of Oxford and Chancellor's Professor at University of California, Los Angeles. She is also affiliated with the Alan Turing Institute and the Farr Institute of Health Informatics Research. Her main research interest is on machine learning and artificial intelligence for medicine. She is an IEEE Fellow (2009) and has been a Distinguished Lecturer of the Communications Society, the Editor in Chief of IEEE Transactions on

Multimedia, and member of the Senior Editorial Board member of IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS) and IEEE Journal on Selected Topics in Signal Processing (JSTSP). She received an NSF CAREER Award (2004), the Best Paper Award from IEEE Transactions on Circuits and Systems for Video Technology (2005), the Okawa Foundation Award (2006), the IBM Faculty Award (2005, 2007, 2008), the Most Cited Paper Award from EURASIP: Image Communications Journal (2006), the Gamenets Conference Best Paper Award (2011) and the 2011 IEEE Circuits and Systems Society Darlington Best Paper Award. She holds 33 US patents.

21.2 Talk

> To make great leaps in data science research to make the world better. ATI Mission Statement.

Three components to this talk:

1. Knowledge Extraction;
2. healthcare policy design;
3. Individualised Treatment Effects.

ML can't do medicine, but can provide actionable information. personalised risk scores; personalised treatment plans; data-induced hypotheses etc.

21.2.1 ICU and Critical Care

Personalised Risk Scoring for Critical Care. In the USA alone, 200K hospitalised patients experience cardio-pulmonary arrests; 75% of these die, but 50% could have been saved. This implies 75K unnecessary deaths in hospital. Hence the solution is to admit people to ICU, but which patients? 12 vital signs, 10 lab tests, 8 admission data.

Graph of blood pressure: very static until falls off a cliff (1100 hours into admission to regular ward), and hence ICU admission. Note that the timing of observations is informative of the doctors' clinical judgements about the state of the patient.

$$X \to M(\Gamma) \to Y.$$

X is hidden state, $M(\Gamma)$ is the model, and Y is the observations.

Note that we have comparatively few patients who are deteriorating. Future depends not only on state, but on time in that state: a few seconds of atrial fillibration is very different from an hour. Hence HMM are not adequate.

HASMM Hidden Absorbing Semi-Markov Model is our development. It captures patient heterogeneity. We have true (hidden) states, and multiple (absorbing) critical states. Model hidden state as a continuous-time stochastic process $X(y)$, and transition probabilities depend on sojourn time (no self-transitions). Model the observation times as drawn from a Hawkes process with

intensity modulated by the condition. Sojourn time is a Gamma distribution, semi-Markov transition probabilities with a multinomial logistic. We do off-line learning [interesting issues if on-line affects the doctor whose observation times we are measuring.] Our prognosis is essentially risk scoring. We can't use standard forward-backward algorithms because both states and transition times are hidden. Define a semi-Markov transition kernel:

$$Q_{i,j}(s) = P(X_{n+1} = j, S_n < s | X_n = i).$$

Theorem 7 *These are the unique solution of a system of integral equation.*

So what about inference? Tradition (Cox proportional hazard model) regress a single static event. Hence we need more. Computing the MLE is impossible as states are hidden. Hence we do expectation maximisation. we don't even know the size of the space *a priori*. But it's not too large: a classical medical theory is that there are three stages of deterioration (we don't necessarily find this: sometimes more states).

Variations in sample times are of different utility in different wards: an interesting point for future research. We use the "informative censoring" (we know why the record ends) we can work backwards to do the expectation maximisation

Cohort of 6094 patients. Two years of data collection (mid2013–2015) and one year of live running[1] (2015–16). Ages 18–114. Many different initial diagnoses. 306 of these 6094 were admitted to ICU. We have 120% sensitivity, and we have 100% PPV improvement over the Rothman system for which UCLA was paying \$500K/year. The timing data didn't improve the final prediction much, but greatly improved timeliness of prediction (which is important). `http://medianetlab.ee.ucla.edu/MedAdvance`.

21.2.2 Heart Transplantation

Personal survival predictions via trees of predictors. Ann/Bob two patients. Urgency: how long will be patients survive while waiting? How much benefit (years of life) will this specific heart that has arrived give each? Current risk scores are pretty bad, partly because they are one-size-fits-all, also linear and horizon-independent, whereas long-term survival is different from short-term survival.

We use Cox regression, linear regression and logistic regression. Choice of regression model is driven by interaction of features. We see 0.76 accuracy(?) versus 0.63 for wait-list, and 0.63 versus 0.54 for post-transplant survival at 10 years. Split data into four sets.

D Basic training set

S_1 JHD didn't quite follow, but these are for setting up the tree.

[1] No updating of the model.

S_2

T testing set

This Tops/R system is not a regression tree, as the labels do not become more homogeneous.

21.2.3 Individualised Treatment Effects

The latest miracle in heart transplantation, LVAD implantation, is dangerous, costs $17K, but a $76K price tag over six-years.

Chapter 22

Operator Scaling — Theory and Applications

Courant Lecture XXXII: 6 April 2017. Avi Wigderson (IAS Princeton). "Matrix and Operator Scaling and their many applications" was the actual title on the day. See [GGOW16] for his second lecture (which JHD couldn't go to).

22.1 Introduction

Analysis Nullcone

Computational Complexity Rank of symbolic matrices, identities, lower bounds

Analysis

Repeat n^c times::
normalise rows $L := R(L) \times L$; normalize columns $L := L \times C(L)$.
Test f $C(L) = I$ up to $1/n$: yes L is NC-nonsingular.
where $R(L) = (\sum_i A_i A_i^\dagger)^{-1/2}$, $C(L) = (\sum_i A_i^\dagger A_i)^{-1/2}$.

22.1.1 Quantum Information Theory

$L = (_1, \ldots, A_m($ where A_i are completely positive maps. $L(P) = \sum_i A)iPA_i^\dagger$ Note that P psd implies $L(P)$ is psd. l is *rank-decreasing* if exists P psd such that $rank(L(P)) < rank(P)$.

$Capacity(L) = \inf\{\det(L(P))/\det(P) : P \text{ psd}\}$. Doubly stochastic equivalent to capacity=1; rank-decreasing equivalent to capacity= 0.

[GGOW15] L is *rank-decreasing* iff L is NC-singular. This is non-commutative algebra. The word problem for free skew fields. $X = \{x_1, \ldots, x_n\}$, and F is a commutative field. Hence $F\langle X\rangle$ is NC polynomials. $F\langle(X)\rangle$ is rational expressions. Note that they don't necessarily simplify: $(x + zy^{-1}w)^{-1}$ [Reu96, Figure 3.2] is genuinely nested inversion. But $r(X) = (x+xy^{-1}x)^{-1} = -(x+y)^{-1}+x^{-1}$:

Hua's identity. Is $r(X)$ zero: word problem. [Amitsur1966] says $r(x_1, x_2, \ldots) = 0$ iff $\forall d \forall D_i \in M_d(F) : r(D_1, D_2, \ldots = 0$.

22.1.2 Invariant Theory

G acts on $V = F^k$ and so on $F[z_1, \ldots, z_k]$. $V^G = \{p \in F(z] : p(gZ) = p(Z) \forall g \in G$

1. $G + S_n$ acts on $V = F^n$ by permuting coordinates V^G is elementary symmetric polynomials.
2. $G = SL_n(F)^2$ acts on V $RLC = (RA_1C, \ldots)$.

Polynomial (semi-)invariants $(Z_i)_{j,k}$ are mn^2 commuting variables. $F[Z]^G = \{p : p(RZC) = p(Z)$ for all $Rm, C \in SL_n(F)$. Degree bounds: [Hilbert1890] $d < \infty$ [Popov1981] $d < \exp\exp n$ [Derksen2001] $D \leq \exp(n)$ which gave rise to [GGOW15].

22.1.3 Analysis

Brascamp–Lieb inequalities. generalises Cauchy–Schwartz, Hölder etc.

But note that we can't solve the commutative determinant question in polynomial time, and being able to do so would have amazing consequences.

Q What about things between commutative and free?

A Good question: natural attack route.

Chapter 23

Kovacic Seminar/CUNY

23.1 Amzallag: on Hrushovskii's Algorithm

Definition 5 *Let* $M_\rho = \{P \in C(t)[X] | \deg(P) \leq \rho \wedge P(F) = 0\}$ $G_\rho = \{g \in GL_n(C) | \forall P \in M_\rho, P(Xg) \in M_p\}$.

So a basis for M_ρ gives definition polynomials for G_ρ. As ρ increases, the add equations, and make the group smaller. The true Galois group is contained in every G_ρ.

Theorem 8 ([Fen13]) *Let* $\tilde{d}$ *be a bound for proto-Galois groups. Then ...*

But his $\tilde{d}$ was 7-fold exponential $O\left(n^{n^{n^{n^{n^{n^{n^3}}}}}}\right)$. We claim $O(n^{n^4})$.

Definition 6 *We say that a group* $H' \subseteq GL_n(C)$ *is a* pre-envelope of index at most m *for a group* $H \subset GL_n(C)$ *if*

1. ...
2. $[H : H \cap H'] = [HH' : H'] \leq m$
3. *every unipotent element of* H' *lies in* H°.

[H° *is the identity component,* H' *is just another group.]*

Question: do these always exist?

Proposition 1 *Let* $H \subset GL_n(C)$ *be an algebraic subgroup. The* H *has a pre-envelope of index at most* $J(n) \cdot 3^{n^4}$*, with degrees of the polynomials bounded by* $k(n) = n^{O(n^4)}$.

Proposition 2 *Let* $H \subset GL_n(C)$. *Then there exists a subgroup* $\tilde{H} \leq GL_n(C)$ *bounded by* $k(n)J(n) \cdot 3^{n^4}$ *such that* $\left(\tilde{H}^\circ\right)^t \triangle H^\circ \leq H \leq \tilde{H}$ *where* $k(n) = n^{O(n^4)}$.

Proof: $\tilde{H} = HH'$. $J(n) \leq n! 12^{n^2}$ [Caesar].

Q–Ovchinnikov Can you compute J?

A–MFS It has a computation in terms of indices of subgroups of $GL_n(C)$. Doable in theory.

Q Why is yours better than [Fen13]?

A We use unitary groups rather than all groups. Also Feng uses Gröbner Bases, and produces generators, which might be large [Chi09].

23.2 Panel: Sunday 10 April

23.2.1 Jobs

Where `MathJobs.org` for maths job is 99%.

23.2.2 Collaboration

Good for a junior person to have some "own papers", as well. Only collaborating with supervisor is generally viewed negatively.

Chapter 24

ACM Blockchain webinar

`http://event.on24.com/eventRegistration/console/EventConsoleApollo.jsp?&eventid=1362212` Mueller-Eberstein,M., The Next Radical Internet Transformation: How Blockchain Technology is transforming Business, Governments, Computing and Security models.

24.1 Talk

Claims that Blockchain solves the problem of transferring value from A to B without a trusted third party. Claims that FinTech is the change agent, and banks are at a tipping point. Blockchain $\neq$ BitCoin.

Many commercial or Government "services" are basically a database. Claims Disney has their own, open source, "DragonChain". The whole talk sounds rather like "proof by investment". Tries comparing the slowth of credit cards being definitive (he claims 90 days) with speed of Blockchain.

But [ME17, Slide 50] is pretty good. Discussed the Ethereum/DAO hack [Ano16, But16]. Claims there's a 51% attack, but we now know better [ES13], [KA16, §4.1.4.1].

24.2 Q&A

Q Incentive adoption model?

A We need to think about this. I doubt the inventors envisaged the massive mining in China, Iceland. We could do mining on mobile 'phones etc.

Q You mentioned $1.7T of inefficiency, but what are the hard costs of a Blockchain?

A Cambridge professor has studied "community" block chains.

Q Could a Blockchain be corrupted via a virus.

A In principle, anything can be corrupted if you try hard enough, but look at the track record. No virus attacks since BitCoin started.

Q Is BitCoin environmentally friendly?

A Compared with what? Driving to the bank? Most "mining" takes place with free/surplus energy.

Chapter 25

Phase transitions in Random Constraint Satisfaction Problem

25.1 Abstract

Speaker: Allan Sly, Princeton University. NYU Courant Colloquium 17 April 2017,

25.2 Talk

Examples of CSP:

1. scheduling appointments
2. solving linear equations
3. colouring a graph
4. satisfying a Boolean formula

The last two are NP-complete and play a central role in complexity theory.

But we are interested in the random version. Erdős–Rényi random graph etc. When is there a k-colouring, is there an independent set of size βn.

Random k-SAT: the or[1] of m clauses each with k literals. m/n is the clause density. NAE-SAT where both $\mathbf{x}$ and $\neg\mathbf{x}$ are solutions. Questions:

1. What is the satisfiability threshold?

[1] JHD is pretty sure this is what he said, i.e. DNF. But much of the literature is CNF. Of course, there's duality.

2. Free Energy: how many solutions are there

3. local statistics

4. Algorithmic – not much is known.

Disordered systems such as spin glasses are models of interacting particles. In particular, Replica Symmetry Breaking, and Cavity Method.

Conjecture 1 (Random k-SAT) *the probability that*

[Friedgut1999] proves that the transition sharpens a possibly non-convergent sequence. $k = 1, \alpha_{sat} = 1$. [Goerdt1992]. Upper bound $2^k \ln 2 - (1+\ln s)/2 + \epsilon_k$. Algorithmic $\geq 1.817 \cdot 2^k/k$. nonconstructive methods do better, $2^k \ln 2 - (1 + \ln 2)/2 - \epsilon_k$. So (non-constructively) α_{sat} is known to within something tending to 0.

First moment is $2^n(1-1/2^k))^m = exp(n(\ln 2 + \alpha \log(1-1/2^k)])$. So if there's one solution, there's ϵn unconstrained variables, so $2^{\epsilon n}$ solutions.

For random colourings and NAE-SAT, second moment works for $\alpha - 1 = \alpha_{SAT} = -O(1)$. Some remarkable results for dense graphs.

Consider two solutions to be adjacent, if they are different at 1 (or a few) Booleans.

Q Hamiltonian?

A The solution space starts out as a well-connected cluster, after α_{clust} SOL decomposes into exponentially-many clusters.

$\exp -\#violations$. After α_{cond} a few large clusters.

$$EZ = \sum (\text{cluster soze}) \times E(\#\text{ clusters of this size}]$$

Dominated by clusters of size s, where $\sum'(x) = -1$. In a typical mass, the mass is dominated by a few clusters of the last size with $\sum = 0$. But, what precisely is a cluster?

RSB *1-step replica symmetry breaking* 1-RSB heuristic says that there is no extra structure at the cluster level and decay of correlation. The heuristic says that, if there are Ω clusters, we apply moments to Ω. With Ding&Sun, author has the exact threshold for regular MAX-IND-SET. [Cojaetal2013] best previous results for k-SAT.

Represent clusters as a new spin system on $V(\mathcal{G})$. Start from $\mathbf{x} \in \{+, 1\}^n$, and is a variable can be either, call it f. So map variable to $\{+, -, g\}$. This is locally rigid.

Let P be the space of probability measures on $[0, 1]$. Define a distributional recursion $R_\alpha : P \to P$. There a messy formula, but $(R_\alpha)^l 1_{1/2} \to^{L\to\infty} \mu_\alpha$. Let $\Phi(\alpha)$ be the expected change in $\log \Omega_n$ to $\log \Omega_{n+1}$. Then the 1RSB prediction α_{sat} is the root of $\Phi(\alpha) = 0$. In fact $\frac{1}{n} \log EZ_\lambda$ is the Legendre transformation of $\sum(s)$.

Neighbourhood profile fluctuations. the degree profile of G is $D_G(\mathbf{d})$. Unfortunately $E\Omega$ is dominated by atypical profile $D*$. each clause has a random multiplicative effect on # clusters. But the product of random IID variables is not concentrated around its mean (Jensen).

Our programme for solving this is

1. work with neighbourhoods of depth R ($R \to \infty$)
2. preprocess graph, removing $n\epsilon_{k,R}$ worst variables
3. Fix R-neighbourhood profile $D_R \approx (R_R)^{typ}$.
4. ...

Basically it's second moment method applied to a very complicated random variable.

But what about small k. The physicists claim works down to $k = 3$. When k is large, the degrees are typically very concentrated about their means, but this isn't true for small k, where the marginal distributions are empirically quite spread out. Hence we would need to repeat this 100-page calculations+smarts) for each k.

Chapter 26

The Unpredicted: In Science, Literature and Politics: Simon DeDeo

26.1 Abstract

We are drawn to the new, the unusual, the unexpected: what we could not predict on the basis of what came before. As vast archives of our cultural past and present go online, scientists can now break out of the laboratory to see how novelty, innovation and creativity are both made and received in the real world.

To track these crucial forms of human experience, Simon DeDeo will introduce simple but powerful concepts from information theory, using examples from Jane Austen and Virginia Woolf. Through collaborative case studies ranging from the speeches of the French Revolution and papers in high-energy physics to the online arguments of Wikipedians and Breitbart commenters, he will show how these tools allow us to ask, and answer, two basic questions: Where do new ideas come from? And how do we respond when they arrive?

Simon DeDeo is assistant professor of social and decision sciences at Carnegie Mellon University in Pittsburgh, Pennsylvania, and external professor at the Santa Fe Institute in New Mexico. He runs the institutes Laboratory for Social Minds, whose collaborative work appears in journals ranging from Physical Review to Cognition and PLOS Computational Biology.

26.2 Talk

Chairman Mao "too soon to tell" joke[1]. I used to study astronomy, but I now study people. When psychology started, we acted like the older sciences, put

[1] But see [McG11].

people in labs, and gave them 10-minute problems to solve, for sums of money that would fit in junior faculty start-up grants. But what about things that really matter. Speech/writing are the *only* means of mind-mind information transfer. We are information transfer engines, but, unlike chips in a computer, with different goals. "advertising works by persuading you to reason poorly, and it's very easy to do".

Deductive Logic Known since at least Aristotle

Inductive logic "most birds can fly" $\rightarrow$ "this bird can probably fly". JMK tried to figure this out in his first book (wrongly). See E.T. Jaynes Bayesian Reasoning.

Hence entropy. The parent (advisor) has something in mind, and the child (grad student) has something in mind. Was on Polish Television until the mathematicians joined and did binary chop, and this killed the game. Example $\{1/2, 1/4, 1/4\}$. Then

$$H(p) = -\sum i = 1^N p_i \log_2 p_i$$

as the formula. Entropy of "Pride & Prejudice" is 9.06 bits , and Virginia Woolf "To the Lighthouse" is 0.13.

Then conditional entropy. Conditional entry of P&P, given previous word is 5.41 bits, so mutual information 3.65 bits. VW is only 2.75 bits.

"Small business" is a very clear signal fro republican, "working family" is Democrat. Does textual inference on political manifestos: shows the signal strength of political manifestos over time, in UK and US. Can see Reaganism, Blairism, as low distinction.

Using a tree wrongly (e.g. a different $\{1/2, 1/4, 1/4\}$ takes 1.75 questions. KL-divergence. "mathematicians are very different from scientists". Took String Theory subset of hep-th. Did a "bag of words" model over text. What is Kl over 1 month, 2 months, 3 years. As well as "surprise given the past", we can do "surprise given the future", i.e. transience. The two actually correlate very well. "What's new is quickly forgotten".

93442 stories following the "Sherlock" series. So what's the Kullback-Liebler divergence, and count the "Kudos" (likes). Here, deviation is negative.

Also "Poetry" magazine. The very first one is awful, next one is "J. Alfred Prufrock". 26,212 poems, 72 poems in Norton Anthology. Early radicalism (Pound, Eliot Stevens), the post-war innovation burst, to genre-conforming. Also looks at most edited page in Wikipedia: actually George W Bush. There are KL-spikes, which he can relate to events, like the ToC changing to add "Controversies about". There are two types of reversion-triggered conflicts in Wikipedia: reject/propose and "try it our".

26.3 Q&A

Has a really good slide with thumbnails of his major data slides.

Harris Manipulation, as in BrExit/Cambridge Analytics.

A I'm interested in what people do. But I am analysing Breitbart. People agree for a whole, and then start disagreeing.

Q That Anthology is the most conservative, so using it as a success criterion is utterly bizarre

A Note that the fan fiction distance form the original is large, but they are internally coherent.

Q Unconvinced about "Digital Humanities" — the choice of corpus is key, and easy [to adjust to suit your target].

A Amazing readership in Fan Fiction, and they re-invented peer review as "beta readers".

Q Bob Dylan's Nobel?

A My question is "what does the system care about".

Chapter 27

The Landscape of Some Statistical Learning Problems

27.1 Abstract

Speaker: Andrea Montanari (Stanford) Title: The Landscape of Some Statistical Learning Problems Abstract: Most high-dimensional estimation and prediction methods propose to minimize a cost function (empirical risk) that is written as a sum of losses associated to each data point (each example). Studying the landscape of the empirical risk is useful to understand the computational complexity of these statistical problems. I will discuss some generic features that can be used to prove that the global minimizer can be computed efficiently even if the loss in non-convex. A different mechanism arises in some rank-constrained semidefinite programming problems. In this case, optimization algorithms can only be guaranteed to produce an (approximate) local optimum, but all local optima are close in value to the global optimum. Finally I will contrast these with problems in which the effects of non-convexity are more dramatic. [Based on joint work with Yu Bai, Song Mei, Theodor Misiakiewicz and Roberto Oliveira] More info: `https://mathsanddatanyu.github.io/website/seminar/#montanari`

27.2 Talk

Given $\{z_1, \ldots, z_n\} \subset \mathbf{R}^d$ want to compute a parametric model $p(\theta)$: *theta* $\in \mathbf{R}^p$.

Definition 7 (Empirical Risk) *minimise* $\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta; z_i)$ *where* $l : \mathbf{R}^P \times \mathbf{R}^d$ *is the loss function.*

Logistic model:$P_\theta(Y_1 = 1|X_i = x_i) = \frac{e^{\langle \theta, x_i \rangle}}{1+e^{\cdots r}}$.

Question 1 (Statistical) *How close is $\hat{R}$ to R*

Question 2 (Computational) *How?*

27.3 Convexity

Examples: logistic regression, robust regression.

27.4 Without Convexity

Example 17 (Binary classification) $z_i = (y_i, x_i) \in \{0,1\} \times \mathbf{R}^d$. $\sigma(u) = \frac{1}{1+e^{-u}}$.*b* $\hat{R}_n(\theta) = \frac{1}{n}\sum(y_i - \sigma(\lambda\theta, x_i\rangle))^2$. *[Rosenblatt1958]: perceptron.*

Example 18 (Mixture of Gaussians) $\hat{R}_n(\theta) = -\frac{1}{n}\sum \log(\phi(z_i - \theta 1) + \phi(z_i - \theta_2))$.

Theorem 9 (Vapnik,...) *Unider suitable conditons [omitted] whp ...*

27.5 Example

Start from random initialiisation, run gradient descent, compute variance of $\hat{\theta}(y)$.

If $n \geq Cp\log p$ then with probability at least $1-\delta$ then various sups are bounded by $c\sqrt{\frac{Cp\log p}{n}}$. He + coworkers have a theorem with the same sort of bounds.

Shows a diagram of population risk versus an empirical risk, where topological degrees differ.

Algorithms:1 — use uniform convergence ....

Theorem 10 (M+) *Assume X_i to be centered sub-Gaussian with $E(X, X^T) \geq \delta I_d$ ...*

Intuition is that the population risk is bowl-shaped. Experimental evidence shows that, as d increases we get a more and more sudden transition from $p(nice) \approx 0$ to $p(nice) \approx 1$ when the correct function of d, d is 1.5.

Example 19 (One-bit compressed sensing) *Non-convex reconstruction.*

Theorem 11 (M+) *Assume X_i to be centred sub-Gaussian with $||X||_\infty \leq Ca$ and*

If $n \geq CS - O(\log d)^d$ then with high probability

1. *the nonconvex problem has a unique local minimiser $\hat{\theta}_n$*

2. $||\hat{\theta}_n - \theta_0||$

Example 20 (Spiked Tensor Model) *For any* $\epsilon > 0$*, there are constants* λ_{1I}*,* $\lambda_{ML}(\epsilon)$*,* $C(\epsilon)$ *such that*

- *if* $\lambda > \lambda_{ML}(\epsilon)$*, then* $E(|\langle \hat{\theta}^{ML} - \theta_0 \rangle) \geq 1 - \delta$
- ...

27.6 ±Convexity

Convexity has global optimum, ..., but nonconvex has ....

Example 21 *Maximize* $\langle x, Ax \rangle$ *subject to* $x \in \{1, 0, \}$.

Also SDP, and rank-constrained SDP. Then we work on a manifold which is n copies of S^{k-1}. For $k \geq \sqrt{2n}$ there are no local maxima.

Partial views on this problem: SDP hierarchy, statistical physics.

27.7 Q&A

Q Is your theorem a worst-case result?

A I'm not sure, but there are ...

Q In the case of a convex loss function, would you change A3?

A Convex is very different.

Chapter 28

Theory and Algorithms for Dynamic and Adaptive Online Learning: Scott Yang

28.1 Abstract

Online learning is a powerful and flexible model for sequential prediction. In this model, algorithms process one sample at a time with an update per iteration that is often computationally cheap and simple to implement. As a result, online learning algorithms have become an attractive solution for modern machine learning applications with very large data sets. However, the classical benchmark adopted in online learning, worst-case guarantees for external regret, may be ill-suited for non-stationary data. In this talk, we will analyze online learning in dynamic environments. We first present a general algorithm for online learning against weighted dynamic sequences with a desirable statistical guarantee. Then, using tools from weighted automata theory, we present several techniques for designing computationally efficient algorithms. Specifically, by representing the competitor class of sequences as a weighted automaton, we introduce methods for automata approximation and compression that can lead to algorithms with drastically improved computational efficiency (in some cases exponential). In the process, we also recover and discover new ways of interpreting existing state-of-the-art methods.

28.2 Intro

Online learning is a game between the algorithm and an adversary

- learners actions Σ
- for $t := 1 : T$
    1. Algorithm picks a distribution over actions
    2. ...
- ...
- External regret $Reg_T(A) = \sum_{i=1}^{T} E_{x_t \sim p_t} \ldots$, essentially difference between online and advance knowledge.

No online learning talk would be complete without a discussion of ads and recommendations.

Ideas:

1. prediction with expert advice
2. online convex optimisation
3. special loss functions (bandit feedback, stochastic losses, strongly convex, exp-concave)
4. connection to other fields.
5. What if no static action in Σ performs well
6. can we design robust algorithms
7. can we design algorithms that do better with easier data
8. structural ensemble methods

28.3 Expert learning with automata

Trivial example where action 1 is poor the first half, second action is poor the second half, then external regret is a bad metric, since we're not comparing with a switcher. [HerbsterWarmuth1998] has tracking regret is a better benchmark. Then adaptive regret [LittlestoneWarmuth1994]. Expert HMMs.

Weighted finite automata (usual state transition graph + weights). Bigram model does really well with Error is independent of T.

$D_1(p$
$q) \leq D\infty(p||q)$. In general bounding D_1 doesn't tell us anything more about D_∞. Really bad example. For model selection, we want to balance the complexity of the hypothesis class with minimizing the empirical risk. Has an algorithm to find a good one within a given computational budget.

28.4 Automaton compression

Can I then compress the automaton. This can be dome exploiting symmetries. Use the1974 idea of failure transitions (labelled with an extra symbol ϕ, essentially "otherwise").

Example where state 0-3 are identical, but each with several (identical) outputs. Replace these by all doing failure to a single new state, which then makes the real transitions. Apparently does well in compressing k-shifting automata. For example the MLE bigram model is $O(N^2Y)$, and this operation reduces it to $O(NT)$. The result is computationally equivalent to Fixed-Share algorithm.

28.5 Future work

What about independence of T. Approximate $\mathcal{C}$ instead of $\mathcal{C}_T$, but it might not be a probability space. Can we look at stochastic losses rather than "worst case" adversarial losses.

28.6 Q&A

Q k-shifting in reality?

A Weather-prediction, when e.g. one forecaster is good in summer, another in winter. Someone else has done this is practice for weather and it happens.

Q So when is k-shifting a better idea than regret?

A Discussion.

Q This reduces rich competitor classes to sequence modelling (of best experts).

Chapter 29

The great subway ride of 1967

29.1 Talk

The compere was the Manhattan Borough Historian.

Peter Samson Grew up in Lovell, Mass., with a train set in the attic. Occasionally went to Boston by train, so fascinating by trains. As an MIT student, occasionally when to MIT. But also tech. Club model railway club.

George Mitchell Peter earned the keys to the club room, which required 40 hours on volunteer work, in a weekend!

Q Didn't your 'tribe' invent "hacker"?

PS Yes, as a positive word in those days.

Q Recommends book *Hackers*.

PS There was a story about an unnamed Flushing youth who had ridden the whole subway system on one token, 25h36m. I had estimates for station-station journey time.

Q Picture of MIT's PDP-6. Weights 1800lbs, printer alone 180lbs etc.

GM I'd met PS at the student radio station, but he'd introduced me to the computer. I had family in NY, so I got to ride the subway at Christmas.

We started at Pacific Park in Brooklyn, since I guy lived there, but he didn't show up! The estimates turned out not to be too bad. We also talked to Don Harold: who became the godfather of the Transit Museum. He told the Transit Police, and the press. A couple had to pay a second fare because of ?? a

terminus where you needed to exit and return. The late Richard Guren kept a log. Sample page displayed.

PS As we pulled into Pelham Bay, the platform was crowded with journalists and photographers, so the motorman had to go slowly, which cost us.

Q 236.7 miles, 77 trains, 25h57m20s.

PS But a guy from Harvard (Geoffrey Arnold) has done it in 23h56m. After that, we got together and wrote a set of rules.

A every mile, which is what we did.

B every station (?) JHD asked later. Be able to get off at every station (but not necessarily every line at every station such as Times Square).

C Passing through every station. This allows expresses.

A Guinness allow you to use more than one fare, as long as you don't use motorised transit. (means you can leap from one branch to another in the Bronx).

PS "That's not the way we did it"! General applause.

PS We then got permission to use the PDP-6 remotely in real time, to replan the route. The actual trip-runners have assistants who broke off and used payphones to advise the control centre of changes. Had we been at MIT, this would have been a long distance call, so we persuaded the MIT amulni association in NY to let us use a room there, with a bank of phones, teleprinter. This needs a team near payphones waiting for us to call. We also had to keep moving these people round the system.

GM Also grateful to people who typed in the paper timetables from the transit authority.

A There was a nine-page operations manual, issued to everyone. Station schematics down to the stairway level of each transfer station.

PS It worked, one computer crash. 25h50m30s.

A The second one didn't get as much press.

PS I've been here, and seem to hit a problem every day. We hit the tail end of a reliable system.

GM The new trains are wonderful to ride (he referred to the old names: IRD, IRT etc.[1]).

A Shows an archival PDP-6 circuit board, which PS got at de-commissioning.

[1] Later, PS said "A BMT station just reeks of BMT!

PS We had no vision in the 1960s that computers would get smaller: for us bigger meant faster.

GM Bull session: "Do you think anyone will ever put a computer in a briefcase" — "Nah, where would you put the CRT?".

PS I'd advise anyone to try it: you may not set a record, but you'll master a complex system.

Showed a five-minute BW film of the run. The [in those days] M train round the Nassau loop only ran three times a day, but we had to catch it as it has a unique piece of right-of-way.

Control centre has a large map, showing all the message-runners, as well as the actual party.

29.2 Q&A

Q What was it like on the subway when no-one was talking on mobiles?

A We had very little interaction. New Yorkers aren't curious.

Q Always in same car?

A Yes, always head car (and we noted the number, so that the log could be verified)

Q Any other cities? What would it be like today?

A London (which is how Guiness got into it). Given the changes, probably 22-23 hours.

Q How close were the data to reality.

A The data helped, but probably not a lot. Enough, though.

Q How did Geoffrey Arnold do it?

A We don't really know. He didn't have a log.

Q Did you scout out the transfers?

A Yes, and a couple of stops before a transfer, we would change to the optimal car for the transfer.

Q Did you do any retrospective analysis with schedules?

A No?

Q Why the formal clothes?

A They weren't formal. I would have sworn I wasn't wearing a tie, but apparently I was.

Q Of those 25 hours, how much was riding?

A About 19.5.

Q When we did it (round of applause for people doing it) we had a sign saying how long we'd been riding it. Also, since I plan these, how do you cope with engineering works?

A Class A is probanly impossible these days.

Q Did you think of doing this afte rthe Cristie street connection?

A No - I declared myself out of competition and on the jury.

Q Must picturesque line.

A Rockaway.

GM Sunrise at Astoria Boulevard.

Q Who would play you in the movie?

A Someone who doesn't want a reputation.

Q When did you do it?

A It was over 24 hours, so we started at a rush hour, to get three of these? Time of year was "not winter".

Q Did you take the same train more than once?

A Yes: express/local.

Q Where's the operations manual?

PS On my website: `www.gricer.com` [British for a railway enthusiast, apparently]

Q Dangerous areas?

GM We were young and invulnerable.

PS We were in the system, which is OK.

Q All men.

A 1 dorm for coeds, seven for men in those days.

Q Bathrooms?

A We knew where they were (in those days there were restrooms in the subway). But we didn't drink a lot.

Q I am the most recent holder of Guinness. That's very different rules: I had to go to Times Square four times.

Q I used machine code on 7090 at those times. Where we will be in 50 years?

A Elon Musk is proposing the merger of humans and machines.

Q What contributes to your low opinion of BMT.

A I don't have a low opinion, but it is distinctive.

Q Did Don Harold make any special arrangements.

A He didn't hold trains: that would be wrong. But he was very helpful.

Q Why 1967 (again).

A A run on a firm set of rules.

Q I noticed the PDP-6 program was LISP: was there any choice?

A Not in my mind (applause from the audience).

Q Did anything surprise you?

A Some transfers, but nothing that really threw us.

Q Monday morning quarterbacking?

A I did play with the computers route, for my own entertainment.

Compere I have owned a subway car, and presented souvenirs to PS/GM.

Chapter 30

Preventing Information Leaks with Policy-Agnostic Programming

ACM Webinar

30.1 Abstract

Most will agree that information leaks are becoming increasingly prevalent. What people may not know, however, is that many leaks are entirely preventable. In this webinar, we explain what it means to be an information flow leak, discuss challenges in information flow security, and convince you that the solution is in using secure-by-construction programming models. More specifically, we will convince you that the solution is the policy-agnostic programming model, where the machine becomes responsible for implementing security and privacy policies with respect to policy-agnostic programs. Using this model for information flow security, a program needs to implement each information flow policy only once, instead of as repeated access checks across the code base. We formally explain what it means for a program to be policy-agnostic, as well as the security guarantees that we have proven. We present both static and dynamic solutions, and extensions of each for database-backed applications. We discuss results showing that, compared to traditional programs, policy-agnostic programs have 1) a smaller trusted computing base, 2) fewer lines of policy code, and 3) reasonable, often negligible, additional overheads (and no overheads with the repair-based solution).

In this webinar, James Mickens and Jean Yang convey this information by playing an interviewer and interviewee for a reality show.

Speakers

SPEAKER Jean Yang Assistant Professor at Carnegie Mellon University

Jean Yang is an assistant professor in the Computer Science Department at Carnegie Mellon University. She received her A.B. from Harvard and Ph.D. from MIT. Her research mission is to develop programming models and tools towards making provable guarantees ubiquitous. During her Ph.D. she created a programming language, Jeeves, that factors information flow checks out of the rest of the program. Her paper on Verve, and operating system verified for type safety, received Best Paper Award at PLDI 2010. Jean also works on analysis tools for modeling intracellular signalling using rule-based graph-rewrite programs. Jean has been studying humorous communication of scientific ideas under the patient tutelage of James Mickens since 2009.

MODERATOR James Mickens Associate Professor at Harvard University James Mickens is an associate professor of computer science at Harvard University. His research focuses on the performance, security, and robustness of large-scale distributed web services. Mickens received a B.S. degree in computer science from the Georgia Institute of Technology in 2001. In 2008, he received a Ph.D. in computer science from the University of Michigan. Before coming to Harvard, he spent six years as a researcher at Microsoft. He is also the creator of Mickens-do, a martial art so deadly that he refuses to teach it to anyone (including himself).

30.2 Non-notes

JHD heard the seminar, but it was indescribable.

Chapter 31

Combinatorial Inference

31.1 Abstract

We are excited to have Han Liu from Princeton University as our speaker this week.

Abstract: We propose a new family of combinatorial inference problems for graphical models. Unlike classical statistical inference where the main interest is point estimation or parameter testing of Euclidean parameters, combinatorial inference aims at testing the global structure of the underlying graph. Examples include testing the graph connectivity, the presence of a cycle of certain size, or the maximum degree of the graph. To begin with, we develop a unified theory for the fundamental limits of a large family of combinatorial inference problems. We propose new structural packing entropies to characterize how the complexity of combinatorial graph structures impacts the corresponding minimax lower bounds. On the other hand, we propose a family of practical structural testing algorithms to match the obtained lower bounds. We use a case study of brain network analysis to illustrate the usefulness of these proposed methods. `www.princeton.edu/~hanliu`

31.2

$x_1, \ldots \sim P$ and $\theta = T(P)$: classical inference has θ Euclidean, hence statistics. If θ is discrete, we have combinatorial inference. So how do we develop valid tests and confidence sets over a discrete structure. So two variables with no line between them are independent, so Markov property. Cpnnectivity, average degree etc.

There's ly one nullgroups. $X \sim N(0, theta)$. $H0 : G \in G_0$ versus $C \notin G_0$. Use symmetric risk: p(H_0error)+p(H_1 error). Use Null-alternative separator, e.g. G_0 disconnected, but adding an edge connects it $C(G_0) = \{e : G_0 \cup \{e\} \notin \mathcal{G}_0\}$. Let this be $\{e_1, \ldots, e_m\}$. This is he single-edge version: we may have

multiple-edge versions. Le Cann's Lemma and Chi-square divergence

$$R(S_0(\theta), S_1(\theta)) \geq 1 - \frac{1}{2}\sqrt{D_{\chi^2}\left(\frac{1}{m}\sum_{i=1}^{m} P_{\Theta_i}, P_{\Theta_0}\right)}.$$

Use graph distances rather than Euclidean distances

Theorem 12 *NeykovLuLiu2016 For any combinatorial test, given any $G_0 \in \mathcal{G}_0$ and its null-alternative separator C, if*

$$\theta \leq \kappa\sqrt{\frac{\log N(C, d_{G_0}, \log(|C|))}{n}}$$

and $N(\ldots) \to \infty$ and $n \to \infty$ then $\exists\kappa$ such that $\liminf_{n\to\infty} R(S_0(\theta), S_1)\theta)) = 1$.

Combinatorial equivalent of [YangBarron19999].

Example 22 *Two circles each with $d/2$ edges, so $(d/2)^2$ connecting options. Plugging this in says it works for $\theta \leq \kappa\sqrt{\frac{1}{n}\log d}$.*

This technique gives us lower bounds for many combinatorial inference problem: acyclic/not; triangle-free/triangle; max degree $\leq s$ versus not etc. But all these work if we have a minimum signal strength condition.

31.2.1 Upper bounds

Null-alternative witness method. Biologists for example really needed a specific bound on Type 1 error. Use the first half of the data for witness searching, an dthe second half for certification. Use $\mathcal{G}_0$ disconnected as example. Then out witness searching looks for a spanning tree $\widehat{W}$. Then the test is that $\widehat{W} \not\subseteq G$. This is a very generic methodology, provided the proerty is monotone (preserved under edge deletion), e.g. disconnected, max degree $\leq s$.

1. Critical witness space is $W(\mathcal{G}_0) = \{G \notin \mathcal{G}_0 | G' \subset G \Rightarrow G' \in \mathcal{G}_0\}$. So disconnect graph/spanning trees, max deg $\leq 4/5$-stars, acyclic/loops. Then witness searching for $N(0, \Theta^{-1})$.
   (a) Estimate $\widehat{\Theta}$ on D_1, greedily adding edges by $|\widehat{\Theta}_{jk}|$'s
   (b) Find the first subgraph that appears in $W(\mathcal{G}_0)$ and output it as $\widehat{W}$.
      * The hardness of this part depends on $\mathcal{G}_0$: easy for disconnected/spanning trees for example
2. Then verification
   (a) post regularization inference to debias
   (b) apply multiplier bootstrap to reject all edges e satisfying

$$|\widehat{\Theta}_e^d| \geq (1-\alpha) - quantileof \max_{e\in\widehat{W}} |\widehat{\Theta}_e^d - \Theta_e|$$

(c) Reject H_0 if alledes in $\widehat{W}$ are rejected.

[Chern. . . 2013]

Theorem 13 *NeykovLuLiu2016 To test any montone property under Gaussina graphical model* $N(0(\Theta^{-1})$*, let*

$$\theta_{\min} \ldots$$

Note that there can be several alternative witness structures.

What about the Ising model rather than the Gaussian model? [NeykovLiu2017].

Chapter 32

East Coast Computer Algebra Day

Held at Wolfram Research, Champaign, Illinois.

Welcome by Roger ??. CA is important, but we're realising that we're still only scratching the surface. Lots of 'hybrid' methods that need to be explored. High complexity issues in Image Recognitoin and ML, and CA hasn't gone that way (yet?).

32.1 Verifying the results of integer progrmaming solvers: Dan Steffy

With Cheung (Carleton) and Gleixner (Zuse).

32.1.1 Background

Problem 3 *Maximise $c^T x$ with $AX \leq b$ and $x \in \mathbf{R}^{N-1} \times \mathbf{Z}^{n_2}$.*

We want "as simple a proof certificate as possible", such that, currently, a human can believe the certificate checker, and eventually we can think of formal verification. Many applications, including union-closed conjecture [Fra95].

There are many solvers, mostly numerical, i.e. $\mathbf{R}_{IEEE}$. Simplex method for the real case. SCIP is state-of-the-art open source: 800KLoC.

1. Then cutting plane method is used to discard a fractional solution $x_i = r$ that simplex finds. So how does one generate a cutting plane?
2. Alternatively "branch and bound": either $x_i \leq \lfloor r \rfloor$ or $x_i \geq \lceil r \rceil$.

32.1.2 Verification

Known techniques for simplex such as incremental precision boosting, knowing which vertext, we can find in **Q**, etc. There's application of directed rounding to cutting planes, generally unpublished. For B&B, there's "safe dual bounds".

For MIP there is a dual problem, but it's no longer a MIP (unlike LP). Note that verifying feasibility for a MIP is easy, it's optimality that's difficult. Note that TSP can be cast as MIP, and this does have verification ideas, verifying the entire algorithm. [Carretal] tried to do the same for MIP. Also tools coming out of Flyspeck. [Applegateetal2009] solved an 85K city TSP. Certificant includes entire branch-bound tree, with dual solutions at each node. In fact, the validity checking includes, recursively, some TSPs. Hence the checker is quite complex. `www.math.uwaterloo.ca/tsp/pla85900`.

Our rule are

1. linear inequality inference (algebra over **R**).
2. Then there's rounding: $y \geq= 1/4 \wedge y \in \mathbf{Z} \Rightarrow y \geq 0$.
3. "Unsplitting" (basically case analysis, matching "branch").

Unsplitting+rounding generates what is known as "split cuts", which are a very general form of cutting planes (general MIP literature).

Our checker incorporates a rational version of SCIP. We try to prune our certificates. This gives us our VIPR.

32.1.3 Computational Experience

Various data, showing verification is about 1/10 of generation time. Average on problems: 22 sec for SCIP exact, 40 for SCIP exact+certficate, 2.2 for certificate tidying, 5.3 for verification.

32.1.4 Conclusions

JHD Straight SCIP?

A Clearly much less, but not listed.

Q–EK We found that interactivity with the solver is key, when doing sums-of-squares?

A Good point.

Q-Roche Same question: challenge/response questions. Can replace challenges with hashes.

A Thanks.

32.2 Wolfram Language

32.2.1 Devendra Kapadia: Introduction

I do probabilistic computations, which had applications for insurance etc. Then also did calculus.

32.2.2 Adam Strzeboński

Radical representation: quadratics are fine. In principle can do 3,4 and some more, but this was lies chaos. Complex numbers even if answer is real, automatic siplification doesn't give 0, etc. Attempt to use further gives 1.5M leaf tree.

Alternative a `Root[minpoly,id]` representation, ordered in $\mathbf{R}$, and over $\mathbf{C}$ by lexicographic ordering of the isolating interval for the real part (N.B., not the same as lex of real parts). Simplifies much faster, and now 21K leaf tree.

There's an `AlgebraicNumber` object in Mathematica. Supports development to any precision?

But what about non-algebraic? Can do similar for exp-log at least.

BDS Parallelism?

A The internals do, and no top-level algorithms don't have to do so explicitly.

32.2.3 Jose Martin-Garcia

This works for any rank, not just matrices.

Dense Arrays Obvious

Sparse Arrays

Structured Arrays e.g. symmetry (more powerful in higher rank).

Symbolic Arrays Can use the assumptions framework of Wolfram language. Build with `Inactive[Table]` which can therefore have symbolic dimensions.

So what operations are possible? Shows operations on symmetric tensors, both actual and symbolic. The product preserves the symmetry. Use double-coset enumeration, which is in theory expensive, but in practice solves the problems physicists ask very quickly. Shows differentiating symbolic matrices (therefore adding a rank).

32.2.4 Charles Pooh: Symbolic Geometry

Live in $\mathbf{R}^n$. `Ball[c,r]` is a typical example. Also `Mesh`. `ImplicitRegion`, `ParametricRegion`. Operations such as `Volume` and various moment, integration tasks. Then there are (Boolean) derived regions.

More importantly, wnat to *use* these; visualise over a region, integrate (numerically and symbolically), PDE-solve.

Q Change of variables?

A Yes, including Stokes' and Green's theorems.

Shows example of maximising with region constraints. Also equation solving.

In general, this is a big project, in which symbolic is a small, but vital part.

32.2.5 Devendra Kapadia: Differential Equations

Hybid DE Example: ball bouncing on steps.

Partial DE Example: Dirichlet (piecewise) problem on Laplace's equation. Gives an inactive sum, which can be evaluated to 300 terms, say. Same imput to `DSolve` and `NDSolve` (numeric).

Sturm–Liouville Problems Asked for a long time. Can now handle (again `DSolve`). Example, where textbook says $\tan\lambda = \lambda$, and he can get the above `Root` (Section 32.2.2) objects.

Also work `MeijerGReduce` so convert Bessel functions etc. This basically inverts what was knows to convert MeijerG into better-known forms. This form is the basis of all Mellin transform etc. work in V11. Hankel transforms then falls out. V11.1 extends to Fourier transforms. Then (inverse) Radon transforms, with their many applications.

32.3 Rainy Day Lemmas #12, 35: Reznick

I'm a consumer.

Conjecture 2 (Shapiro) *Every polynomial in H_{de} can be written as a sub of at most d dth powers of forms in H_e.*

$d = 1$ is vacuous. $d = 2$ can be done.

New work $(d, e) = (3, 2)$.

$$d^2 p = \sum \zeta_d^{-k} \cdots .$$

Theorem 14 *Given a polynomial $F : \mathbf{C}^n \to \mathbf{C}^m$ Then either*

1. *The polynomials are dependent and the image is a manifold*
2. *The polynomials are independent and the image is the complement of a manifold.*

There are 15 coefficients in a ternary quartic. So what can't it be a sum of five ternary fourth powers? Clebsch's disproof by singularity theory. Sylvester's proofs by catalecticants (he apologised for not using the correct phrase meicatalecticants — same paper as "unimodular").

Lasker–Wakeford comes from H.W. Turnbull, who was the last pre-Hilbert invariant theorist. See his 1960 book (Dover).

Theorem 15 (LW) *If $F : \mathbf{C}^N \to H_d(\mathbf{C}^n)$ then F is canonical form iff ....*

Cayley and Sylvester were both actuaries at the same firm of lawyers in London.

Theorem 16 (Sylvester) *$p(x,y) = \sum_k \lambda_k(\alpha_k x + \beta_k y)^d$ iff the Hankel matrix of coefficients of p deflated by binomial coefficients has a nontrivial null space.*

It is easy to see that if $m < r$, then any polynomial map $F : \mathbf{C}^m \to \mathbf{C}^r$ mushave the property that $\{f_j(t_1, \ldots, t_m)\}$ must be algebraically dependent (but proved by Hilbert, by dimension counting).

Theorem 17 (BR) *Suppose p is a binary sextic. p is the sum of two cubes iff*

1. *p is a perfect cube, or $p = f_1 f_2 f_3$ where the f_I are linearly dependent*
2. *case by simultaneous diagonalisation.*

Example 23 $(x^7 - y^7)/(x - y) = x^6 + \cdots +$

Can you write a real sextic as a sum of cubes three real quadatics? We don't know. Writing octics as sum of sourth powers of quadratics? My co-authors claim this.

Q–EK Have you thought of multipliers and denominators?

A No. Real/complex distinctions. Complex ones would have roots and hence singularities and not universal.

32.4 Take it to the limit, one more time: John D'Angelo

Consider $f(x,y) = fracx^2yx^2 + y^2$: in $\mathbf{C}^2$ this is undefined on two lines, but in $\mathbf{R}^2$ we have a removable singularity.

I'm in my office many years ago. Dan Grayson says "Freshman calculus: $\frac{0}{0}$", L'Hopital, still $\frac{0}{0}$, ..., when do we stop? Answer: series calculations for numerator and denominator.

Voltaire: "If you want to talk to me, define your terms". Hence ϵ/δ definition. Homogeneous function is positive definite if $\exists C > 0$, such that $||f(\mathbf{x})|| > C||\mathbf{x}||^m$. m is even.

$f(x,y) = \frac{x^2y^3}{x^4+y^6}$. Might guess that answer is 0, by weights, but in fact on ...

$g9x,y) = x^4 - ax^2y^2 + y^4$. Then g is positive way form the origianl iff $a < 2$.

$f(x,y) = \frac{x^a y^b}{x^4+y^4-x^2t^2}$. Wolfram gets this one right.

There's a case which hangs. $f(x(t), y(t)) = \frac{t^{2m+2n+\cdots}}{t^{\cdots}}$, ...

1. given $f = \frac{p}{q}$
2. when power series, expand in homogeneous parts

3. order of vanishng of p less than that of q, no limit
4. otherwise ask when the lowest order part of q is ositive definite. If so, easy
5. Look for a substitution for definiteness
6. Try Lagrange multipliers
7. Try pulling back to arbitrary curves
8. Ask a good analyst.

Q–EK Why wrong?

A–DL In the Alpha code we convert to polars, and .... It's "a

$$\frac{\text{obviously zero}}{\text{doesn't look like 0 but is}}$$

becoming zero too early" case.

32.5 Berezinians and Darboux Transformations on the superline: Shemyakova

Introduced by JHD.

1505.05194 and 1605.07286. [LS16].

Describing all Darboux transformations on the superline. Solved completely, analogous to classical case, but not easily so. This gives lots of new problems in superlinear algebra.

A DT maps a differential operator to another operators "of the same form" together witha linear transformations between kernels or arbitrary eigenspaces. Popular in 19th century, then forgotten until 1970s, thenemerged in [WahlquistEstabrook1973]. [Matveev1979] introduced the term "Darboux transformations" and devleoped in integrable systems theory.

Supergeometry. The superline is a $1|1$-dimensional supermanifold with one even coordinate x and one odd coordinate ξ. Everything has a parity, which combined on multiplication.

$$\underbrace{f(x,\xi)}_{even} = \underbrace{f^0(x)}_{even} + \xi \underbrace{f^1(x)}_{odd}$$

etc., so a second order operator is

$$L = a_{20}\partial_x^2 + a_{11}\partial_x\partial_\xi + a_{02}\partial_\xi^2.$$

Everything can be written in terms of[1] $D = \partial_\xi + \xi\partial_x$, and $D^2 = \partial_x$ because of some cancellation rules ($\xi^2 = 0$). Everything can be written in terms of D. An operator is *non-degenerate* if the leading coeficient is invertible.

There's a ring $DO(1|1)$.

[1] Known as 'superderivative' or 'covariant derivative'.

1. If $D\phi(x,\xi)=0$, the ϕ is constant.
2. If $\partial_x\phi(x,y)=0$, then $\phi(x,\xi)=\phi(\xi)$ only, so is $c_1+\xi c_2$.
3. If $\partial_x i\phi(x,\xi)=0$, then *pphi* is 9any) function of x alone.
4. Every monic first-order operator $M=D+\mu$ has the form $M_\phi = D - D\ln\phi = \phi\circ D\circ\phi^{-1}$. $M_\phi(\phi)=0$.

It is possible to divide an arbitrary N by non-degerate M form the left and right, with remainder.

Lemma 1 (Bézout) *Let L be nondegenerate of order m ...*

A *supermatrix* has rows and colums labelled indicating parities (nothing to do with parities of the entries) An even matrix has even entries in EE and OO places, and odd elsewhere. Standard format means all the E rows/columns come before O rows/columns. Don't change order within even, odd rows/columns.

Berezinians is essentially a generalisation of determinants. Parity reversion changes all the columnn labels. For an invertible A, $\operatorname{Ber} A^\Pi = (\operatorname{Ber} A)^{-1}$. Call this $\operatorname{Ber}^* A$. We have new cofactor expansions of Berezinians. A "wrong" matrix is an even one with one odd row/column. By linearity extend the devinition of Ber to this case.

Ber and Ber^* are invariant under linear transformations, provided we only add "correct" rows to the "wrong" one. $\operatorname{Ber}\begin{pmatrix}\xi & x\\ \xi & x\end{pmatrix} = \frac{\xi - x x^{-1}\xi}{x} = 0$

Can expand Ber along even rows/columns, and Ber^* we can expand along odd rows/columns. A formula for adjoint, but more complication.

$$\operatorname{Ber}\begin{pmatrix}x & \alpha\\ \beta & y\end{pmatrix} = \operatorname{Ber}(y)\cdot x - \frac{\operatorname{Ber}^*(\alpha)}{\cdots}\cdot\alpha = \cdots.$$

WRT D we can define superjets, so for f we have $[f, DF, D^2f, \cdots]$. If f is even, its n-superjet is an even vector of dimension $k=1|k$ if $n=2k$ ....

Define superWronskins in terms of Berezinskians, but the natural way of writing this down is not normal format, so need to swap rows and columns to get standard format.

Theorem 18 *A monic differnetial opeator on thesuperline is deifned by its kernel. If ϕ_i is a basis for $\ker M$, then the action of M on arbirartry odd ϕ is given by $M\phi = \frac{W^*(\phi_0,\ldots,\phi_{n-1},\phi)}{W(\phi_0,\ldots,\phi_{n-1},\phi_n)}$.*

Hence we can define Darboux transformation as $M=M_\phi$ where ϕ is even $\in E_\lambda * L_0$.

Theorem 19 *Every DT can be written as a product of first-order ones.*

To every DT we have an invariant subspace of L_0 of a specificed dimension.

If we have DT, we can also define Dressing transformations.

We would like a computer algebra package to work with operators on supergeometric settings.

Q What's the equivalent of characteristic polynomial?

A Good question?

Q Super eigenvalues?

A There's an even/odd distinction here, but not sure of details.

32.6 JHD spoke

32.7 Closing

DL 25 years and one day ago, SW calls us in to look at the threat of Axiom, but I couldn't stay, as my wife had just gone into labour.

JJ Thanks. Aim to have ECCAD 2018 in CY with Andreé Platzer.

Chapter 33

Stochastic Variance Reduction Methods for Policy Evaluation: Lihong Li (MSR)

33.1 Reinforcement learning

Policy $\pi : s \mapsto a$, aiming to maximise long-term reward $V^\pi(s) := E\left[\sum_{t=1}^{\infty} \cdots\right]$. Often a crucial step is to evaluare a fixed policy.

Q Difference between this and Temporal distribution?

A Two names for same thing.

In practice the first two show sublinear convergence and linear-time update, while the other two show linear convergence and linear-time update. Some graphs, but not very easy to read. Benchmakrs are Mountain Car and one other. VRG and SAGA come from convex optimization, which have been extended to saddle-point optimization [BalamurganBach2016]. This needs strongly convex-concave objectives, but we only need strong concavity in the dual.

Conclusion: this is a first-order algorithms with linear convergence rate, and proising experimental results on benchmarks.

Table 33.1: Running times for various algorithms

LSTD	$O(nd^2)$
GTD2	$O(dk_1/\epsilon)$
PDBG	$O(ndk_2 \log \frac{1}{\epsilon})$
SVRG/SAFA	$O(nd1_f rack_3 n) \log \frac{1}{\epsilon}$

33.2 Q&A

Q [Sutton2009] has another algorithm as well. Also you're taking advantage of finite sum idea: infinite?

A See paper for the other algorithms. We can allow infinite time horizon.

Chapter 34

Navigating PU_2 with Golden Gates: Peter Sarnak

34.1 Background

"The final frontier of Physics is Number Theory". Looking at gates in a quantum computer. Classical computer: binary. All circuits can be written in not/and logic. The siz eof a circuit is its complexity.

Theoretical quantum computing. A single qubit state is a unit vector $\phi \in \mathbf{C}^2$. $\psi = (\psi_1, \psi_2)$: $|\psi|^2 = \psi_1\overline{\psi_1} + \psi_2\overline{\psi_2} = 1$. A one-bit quantum gate is $g \in U(2)$ (or $SU(2)$, $PU(2) := G$) acting on ψs. XOR plus all unary gates are sufficient.

Theorem 20 (Solovay–Kitaev) *Given A, B topologically generating G, we can find a word $W(A, B)$ of length $P(log1/\epsilon)^c$ and is as many steps such that $d(w, g) < \epsilon$. Get near randumply, then use the xponentail map and commutators.*

Problem 4 *Given a finite subgroup C of G to find an involution T with $T^2 = 1$ such that $F = C \cup \{T\}$ generates G topologically optimally. A t circuit is $C_1TC_2 \cdots TC_t$. I want uniqueness. Also want them to cover G optimally. $Vol(B)N_F(k) \geq 1$ to cover, but want $Vol(B)N_F(k) \to \infty$ very slowly with k. Also want an efficient algorithm to find the representation.*

5 Platonic solids, but duals have the same group, so only three groups., A_4, S_4 and A_5 of size 12,24,60: all $p + 1$ for p prime. Hence we can produce a group and a T in each of five cases. [ParazanchevskiSarnak].

Algorithm (Ross–Selinger) has an efficient heuristic (assuming fast integer factoring) algorithm to navigate to a diagonal g. Onthe other hand, for abritrary g, then finding the shortest circuit is essentially NP-complete.

What about $U(1)$. $U(1) = \{e^{2\pi i\theta} : 0 \leq \theta < 1\}$. Best is rotating by R_ϕ for $\phi = (\sqrt{5}-1)/2$. Due to Graham/van Lindt/Vera Sós. Use Continued Fractions.

Recall $SU(3)$ is isometric to $S^3 \subset \mathbf{R}^4$. The arithmetic setup is that the words in f of T-count t correspond to solutions in in integers of $x-1^2+x_2^2+x_3^2+x_4^2 = n$,

either in $\mathbf{Z}$ or an ANF. Let there be $N(n)$ such representations. Map these onto unit sphere, and how well do they cover the unit sphere. This is optimal in the previous sense, based on Deligne's proof of Ramanujan's conjectures.

For navigation we should first consider $x_1^2 + x_2^2 = n$ iff $n = \prod p_i^{e_i}$ iff e_i even when $p_i \equiv 3(4)$. Want to solve this in Poly$(\log n)$. Use Schoof's deterministic $O(\log^9 p)$ for prime p. The algorithm in pratcice may give up on factoring and find the second-best solution.

Sppose now given n, α, β and want a soiution with $\alpha \leq x_1/x_2 \leq \beta$. In fact it's NP-complete. Proof sketch: reduce to a subsum problem. This is the obstacle to navigation to non-diagonal problems. I believe that factoring isn't necessarily hard: we haven't any good reductions. Uses the fact that integerporgrmaming in fixed dimension is P (Lenstra). the last step involves factoring $\gamma \in \Gamma = \langle C, T \rangle$ into a word with minimal T-count.

There is an explicit homomorphism $\Gamma \to PGL(2, \mathbf{Q}_p)$, with $p = |C| - 1)$, and such that Γ acts simply transitively on the edges of the $|C|$-regular tree $X = PGL(2, \mathbf{Q}_p)/PGL(2, \mathbf{Z}+p)$. The t-count corresponding to distance moved on three tree.

The miracle of these gates is that this simple transaction only exists for fintely many such Γs, but the right ones exist.

Chapter 35

Vital Math: Math Encounters: Chris Budd

Event at the National Museum of Mathematics.

35.1 Introduction: Bob Kohn

Paul wants to exit, Carol wants to stop him. Paul chooses a direction, Carol may reverse it, then Paul steps ϵ. Difficult for a rectangle. For a circle, Paul points along a tangent, so whether reversed or not he makes progress. Hence circumscribe a circle.

35.2 CJB

Gresham Professor of Geometry, the oldest in the country. At Bath I find linkages with industry.

I'm very positive about the fact my son's a mathematician, but we don't have a great image.

1. Completely useless
2. Mathematicians are evil soulless geeks
3. Mathematicians are mad

And in reality there are problems with writing formulae while waiting for a plane to take off.

A smart 'phone is stuffed full or mathematics. Quote from president of Exxon. Image of a mathematician, turned out to be Maxwell (correctly guessed). CJB is RI Professor, as was Faraday, who discovered the link between electricity and magnetism. axwell turned Faraday's experiments into equations (shown). Note that then can be solved for light, but have other solutions, which turned out

to be radio waves, hence Hertx, and then Marconi. Also Florence Nightingale. Everyone thinks she's a nurse, but she's really a statisician (early member of RSS). Wonderful data presentations.

Maths has changed the world, largely through computers. Many traditional uses, but also computer graphics in the film industry (quoting Shrek). Pixar employs lots of mathematicians. Also *Lord of the Rings*. Consider "people who bought this book" recommendations, also mathematical.

Math was invented to count things with (decimal, or Babylonian by counting on knuckles, etc.). We are pretty sure that the first application was taxation. See Babylonian tablets, Rhind papyrus. Then get quadratics, doubling the area of a field. Babylonian tablet shows 1.4142. Then calculus.

35.3 Mazes and labyrinths

Mintaur. Theseus was Ariadne, first mathematician: gave Theseus a sword and an algorithm. Claims native Americans also had these. From labyrinths to mazes, and Euler's work on mazes and then on networks. COlleague explained how to draw a labyrinth.

35.4 Mathematics and Music

Based on Pythagoras. Note theorem was already known to Chinese. But he did the work on the octave, and musical harmony/fractions correlation. Then the "scale" based on simple fractions. Thne Bach and the ewll-tempered clavier, all based on geometric progressions. As well as music, a big problem of the18th cecntury was navigation, specifically longitude. Latitude solved via the sextant. Longitude depended on the clock, and a mathematics/technology hybrid. Ephemerides, computed by humans (often women) known as computers. Then the midshipmen did a 22-step computation at the end with the observations.

35.5 Applications

scanners have revolutionised medicine. Image of Radon. His problem was reverse engineering an object from its shadows. Realised that Radon's transform could make X-Rays more useful, but needed technology (Cormack, EMI) to actually do it. One application at Bath is scanning of bee-hives. Claims that Killer Sudku is the same maths as scanning.

Images of Saturn, the first one taken by Voyager. beamed to Earth with a 30watt transmitter. Picture truned into numbers, then into a seriously ECC. Image of Hamming. Needed in mobile 'phones etc.

35.6 Q&A

Q Harrison timepiece?

A Quality of bearings, in particular temperature-resistance.

Q Blockchain etc.?

A I've very largely avoided finance, but it's connected to big data and information theory.

Q How can we get your talk at my high school?

A Being done by MoMath, also mail me.

Chapter 36

Murder, Matrices, and Minima - Adventures in Blind Deconvolution: Thomas Strohmer

36.1 Abstract

I will first describe how I once failed to catch a murderer (dubbed the graveyard murderer by the media), because I failed in solving a blind deconvolution problem. Here, blind deconvolution refers to the following problem: Assume we are given a function y which arises as the convolution of two unknown functions g and h. When and how is it possible to recover g and h from the knowledge of y? Blind deconvolution is a topic that pervades many areas of science and technology, including geophysics, astronomy, medical imaging, optics, and communications. Blind deconvolution is obviously ill-posed and even under additional assumptions this is a very difficult non-convex optimization problem which is full of undesirable local minima. I will present a host of new algorithms, accompanied with rigorous theory, that can efficiently solve the blind deconvolution problem in a range of different circumstances. The algorithms will include convex and non-convex algorithms, and even a suprisingly simple linear least squares approach. The mathematical framework behind our algorithms builds on tools from random matrix theory combined with recent advances in convex and non-convex optimization.

Applications in image processing and the future Internet-of-Things will be described. Thus, while the graveyard murderer is still on the loose, recent progress in blind deconvolution may at least have positive impact on the Internet-of-Things.

36.2 The Graveyard Murdered

Austria (home country) 20 years ago. Widow in large house in rich area. Hears burglar, 'phones police who take their time, she goes down, and is killed by the burglar. Police search area in vain. Next day police search graveyard, and find loot. Lie in wait, but fail to ctach burglar. However, police had a camera. But picture is pretty poor (cheap camera). Screwed up three times, so asked a mathematician! We basically failed, wheras Stan Osher/Rodney King case succeeded

$y = g * z + w$ where g is the blurring operator and w the noise. Stan Osher knew the blurring operator, we didn't. This is very ill-posed. Not ethat $z(t) = \overline{g}(-t)$ we have *phase retrieveal problem.*

Otherwise serious uses are IoT. Typically very short packets, hence need to work on overhead challenge.

> Channel estimation in the elephant in the room for 4G — ??? (stanford)

Suppose we have r sensors, each transmitting $z_i * g_i$. How do we recover the z_i, or a wanted specific one.

Consider $\hat{y} = diag(BH_0)\overline{A(X)_0} + w$. Where B is a $L \times N$ low-pass Fourier matrix. A is a Guassian randm matrix and $w \sim N(0, \sigma^2 d_0^2/L)$. If $L \geq C \cdots$ then there is a fast recovery, Four key properties

1. Local Regularity Condition: objective function decreases $||\nabla \tilde{F}(h, x)||^2 \geq \omega$
2.
3. Local smoothness condition
4. Robustness condition

Now about nonconvex versus comvex optmisation. Nonconvex is actually faster and needs fewer measurements.

Blind deconvolution for SIMO OFDM. Typically two receive antennas on a mobile phone. Then this fits into our framework.

In previous examples, the subspaces of the blurring function were known. The theory says that the number of measurements scales like $l = \tilde{\mathcal{O}}(s^2(K+N))$, but numerically not a problem. But in practice linear, and we have a theorem recently.

Blind deconvolution is a special case of self-calibration.

Q Murder?

A The police has done some pre-processing themselves. Also resolution is poor.

Chapter 37

Circuit Complexity: New Techniques and Their Limitations: Aleksandr Golovnev

37.1 Abstract

Candidate: Aleksandr Golovnev Advisors: Yevgeniy Dodis/Oded Regev

Committee: Prof. Yevgeniy Dodis (NYU, CS, advisor, reader) Prof. Oded Regev (NYU, CS, advisor, reader) Prof. Ryan Williams (MIT, CSAIL, reader) Prof. Subhash Khot (NYU, CS, auditor) Prof. Rocco Servedio (Columbia, CS, auditor)

Date/Time: Friday May 5th at 12:00 pm Location: 251 Mercer St., Room 1314 Title: Circuit Complexity: New Techniques and Their Limitations

Abstract:

We study the problem of proving circuit lower bounds. The strongest known lower bound of 3n-o(n) for an explicit function was proved by Blum in 1984. We prove a lower bound of (3+1/86)n-o(n) for affine dispersers for sublinear dimensions.

We introduce the weighted gate elimination method to give an elementary proof of a 3.11n lower bound for quadratic dispersers. (Although currently there are no explicit constructions of such functions.) Also, we develop a general framework which allows us to turn lower bounds proofs into upper bounds for Circuit SAT problems. We also study the limits of the gate elimination method. Finally, we prove strong limitations of the developed techniques.

37.2 Talk

Lower bounds are $3n$, but upper are 2^n, as in [SaxenaSeshadhri2013a] "quite depressing".

Notethat we deal with circuit complexity, which is a tedious area in terms of proofs.

37.2.1 Gate Elimination

Straight-Line Program = Circuit. Gates are binary functions, unlimited fanout, depth. [Shannon1949] Amost all are $\Omega(2^n/n)$ and [Lupanov1958] all can .... In general for explicit n most functions are very close to the bound, but the actualt greatest number is $3n$ if $f(x, a, b, c) = x_a x_b \oplus x_c$. $3n$ also proved by affine dispersers.

3.011n (3+1/86)n for affine dispersers, and better for quadratic dispersers (non-explcit)

1. Show that, for any circuit computing $f \in C$, we can find a substitution eliminating at least three gates
2. Show this function is still in C
3. Induction

There are 16 bnary functions: 2 constants, 4 degenerate, 2 xor-type and 8 and-typo

37.2.2 Affine Dispersers

A function is an affine disperser for dimension d if it is non-contant on any affine subspace of dimension at least d. Equavalently, and $n-d$ linear restrictions don't make it constant.

Also, allow Xor-layered circuits.

For a ciruit computingan affice disperser for dimension d,$inputs(C)+gates(C) \geq 4(n-d)$. Proof by making $n-d$ affine restrictions, each tiemme reduciig inuts+gtes by at least 4. Then convert C to XOR-layered, and take a top gate A. For outdegree 1, set X to 0, and changes gates, inputs by 2. Simialrly.

37.2.3 Our proof

Delayed linear substituions $x_3 := 0$, $x_3 := x_7 \oplus x_{10} \oplus 1$ $x_3 := x_4 x_7$. These are quadratic, but we "promise to eliminate later": $q = \#$ quadratics.

Cyclic circuits Scary! Let b be nuber fo bottleneck gates.

Then $\mu = g + \cdots$ a complexlinear weightings on others.

37.2.4 Quadratic dispersers

Note that a random function is a very good quadratic dispersers. $(n, 2n, , 2^{n/100})$ is all we need. [dimension, #contraints, # nonvanishing set].

Then weighted gate elimination says we reduce S by a factor of α, and be sure to eliminate at leasr $3 \log \alpha$ gates.

37.2.5 Circuit Satisfiability

Algoriths for circuit sat use the same gate analysis. Our framework produced worst-case lower bound for a disperser, average case circuit lower bound for an extraction, Algorithms solving #SAT(C) in 1.99^n.. Gives new bounds [GKST15].

37.2.6 Limits

For all these lower bound functions, we have equavalent upper bunds, also linear, so no nonlinear lower bound. Theroem [GHKK2016] No nonlinear lower bound this way.

Theorem 21 *Ther are functions of any circuit size such that after* any *substitution* $x := \rho$, $gates(f) - games(f|_{x:=\rho}) \leq 5$. *Here* ρ *could be any function of the others.*

Chapter 38

Effective Bounds for Differential Equations: Pogudin

Actually a talk at the Kolchin Seminar.

38.1 Polynomials

Problems: consistency checking of system p_i, elimination into fewer indeterminates. Nullstellensatz says that inconsistency is $1 \equiv 0$. More precisely $\exists q_i : \sum q_i p_i = 1$. But how to search for q_i: need a bounded version. See [Her26]. Elimination also means $\exists q_i : \sum q_i p_i = r(x_1, \ldots, x_s)$.

38.2 ODEs

Could ask for solutoins in power series, or in anaytic functiosn near zero, actually equivalent.

Example 24 *$x(t)x'(t) = 1; x''(t) = 0$. Just adding values for $x(0)$ etc doesn't help, but if we also consider $(x(t)x'(t) - 1)' = 0$ then the polynomial system is inconsistent.*

But how often should I differentiate? Bounds by [Grigoriev1988], also more recent ones with ineffective constants.

Similarly for elimination. Here we don't know N. An alternatve approach is Rosenfeld–Gröbner.

38.3 Our work

With Ovchinnikov, Ngo Thieu Vo have a bound for both the Nullstellensatz and eliminaton.

- polynomial in degrees of equations
- expponential on the number of distinct algebraic variables
- doubly exponential on the dimension of the variety defined by the initial equations and polynomial equations.

\+ for elimination we count only the varibales being eliminated.

Example 25 (SIRS model) *According to his bounds above, need 164 differentiations. But the corresponding ideal is prime, so in fact 33 will do. Knowing that the ideal is of degree 2 reduces to 9. That it's a trinagular set reduces to 5.*

Example 26 (HIV model)

$$T_t = (\rho - k_m T_m - k_w T_w - k_r T_{mw}) \tag{38.1}$$
$$(T_m)_t = \tag{38.2}$$
$$(T_w)_t = \tag{38.3}$$
$$(T_{mw})_t = \tag{38.4}$$
$$\tag{38.5}$$

We can produce formulas that express the values of the parameters in terms T etc. This reduces the number of measurements needed from 20 to 16.

Chapter 39

The Geometry of Similarity Search

39.1 Abstract

How does one efficiently search for similar items in a large dataset, say, of images? This is a classic algorithmic task that arises when dealing with massive datasets. Its applications span many areas, including information retrieval, machine learning, data mining, computer vision, signal processing, bioinformatics, etc. For example, this task underlies the classic 'nearest neighbor' classification rule in machine learning. The natural solution for similarity search - to scan the entire dataset, comparing a given item to each item in the dataset - is prohibitively slow for modern datasets.

Alexandr Andoni will describe how efficient solutions for similarity search benefit from the tools and perspectives of high-dimensional geometry. The latter emerged as the natural language of similarity search: e.g., a 20 x 20 pixel image naturally corresponds to a 400-dimensional vector, one coordinate per pixel. Andoni will show how geometric tools such as dimension reduction and randomized space partitions enable efficient similarity search algorithms.

Dr. Andoni is associate professor of computer science at Columbia University and a member of its Data Science Institute. Previously, Andoni performed research at the Massachusetts Institute of Technology, the Center for Computational Intractability (Princeton University, New York University and the Institute for Advanced Study), Microsoft Research in Silicon Valley, California, and the Simons Institute for the Theory of Computing at the University of California, Berkeley. Andoni focuses on advancing algorithmic foundations of massive data, studying topics such as sublinear algorithms, high-dimensional geometry, metric embeddings and theoretical machine-learning.

39.2 Introduction

This is the third annual meeting of Algorithms and Geometry Simons-funded seminar. This is the public lecture concluding this conference.

39.3 Talk

I'm a compter scientist, so work on algorithms. Given a database of images (from `imagenet`), find similar images. So what is similarity. Naïvely $O(n^2)$. But $n \sim 10^9$, and the images may be on many computers. Images (BW) as 2D array of pxiels. Then think of this as a high-dimensional vectors. Then Hamming distance between two such. Need more subtle tools to deal with rotation, scaling etc. Hence need "feature matching". Then we need "earth-mover distance" to look at similarity.

Definition 8 (Nearest Neihbour Search) *Given a database of a set P of points, and a query point Q, to find the closest message in P to Q. Parameters, n points in dimension d. Offline preprocessing is legal.*

Clustering is a related problem. I encountered this problem for code fragment duplication detection at Microsoft.

Q All points have the same dimensions?

A Yes — an issue we try to hide.

Special case: exact duplicate detection. Also nearest-neighbour, k-nearest neighbour.

Locality-sensitive hashing. [IndykMotani1998]. Code the points to that "similar" becomes "exact". This coding is a partition fo Euclidean space inot regions assigned the same code. Mp g on $\mathbf{R}^d$ such that $||q-p|| \leq r \Rightarrow g(p) = g(q)$ and $||q-p|| > cr \Rightarrow g(q) \neq g(p)$. But of course not possible (for reasonable disance functions).

Hence introduce randomisation: $||q-p|| \leq r \Rightarrow Pr(g(p) = g(q))$ is not too low ($\geq p_1$, etc. ($\leq p_2$). Then use several such indices. Depends only on r and cr. Number of indices is n^ρ where $\rho = \frac{\log(1/p_1)}{\log(1/p_2)}$.

regular grid randomly shifted and rotated. Space $n^{1+\rho}$, time n^ρ, and typical[1] $\rho = 1/c$ with say $c = 1/2$.

ball carving p can hit empty space, so add enough such grids to cover the whole of space (nice image). But needs d^d grids.

Q Why cover the whole of space, not just the points given?

A Good question, but even covering a reasonable fraction is similar.

[1] It seemed that $\rho = 1/c$ always, but JHD didn't see why.

Therefore Do this on projections onto a t-dimensional subspace. Best t gives $\rho \to 1/c^2$.

These space partitions are ubiquitous: graph partitioning, metric embeddings, communications complexity etc.

Can prove [O'DonnellWuZhou2011] that $\rho \geq 1/c^2$ always. This is essentially an isoperimetric problem.

Q Random t-dimensional space? Why not the "best" for the data?

A Good question, but there's not much theory in this area. Also what about worst-case?

39.4 Other LSH

Yes, there are better maps, i fthe maps cn depend on the data.

Example 27 (non-exmaple) *Define $g(q)$ to be the identity of the cloest point to q. It's the ideal map, but computing it is the problem.*

This can get to 1/7 rather than 1/4. Nice configurations (more than general position, want all vectors to be pseudorandom) given better LSH, and there's a reduction that gets any configuration to nice.

Chapter 40

The Great Unknown: Marcus du Sautoy

Museum of Mathematics, June 7th, 2017. Subtitle: Seven journeys to the frontiers of science. Introducer: "The more we know, the greater our perimeter of ignorance becomes".

There is so much science: possibly Newton and Galileo were the last to "know it all". But can this ever be true collectively? Or are there things we can never know? I think these can be called "seven edges".

> Everyone by their nature desires to know — Aristotle.

The track record is bad, e.g.

> We will never be able to study the chemical composition of stars — Auguste Comté.

Rumsfeld: "unknown unknowns". These I can't talk about by definition, so I'll do known unknowns.

40.1

Mathematicians are pattern searchers. If we have the equations, can we predict the future. The die is random, but post-Newton, we should be able to predict. But consider Poincaré. A small error in the present can blow up.

Example 28 (Double pendulum) *Simple equations, but can we predict?*

Example 29 (Three magnets and pendulum) *Again simple equations, but can we predict?*

Example 30 (Casino) *Tossing a coin ten times, can you not get 3 H (or 3 T) in a go. If you get it right, the Casino will pay you \$1...\$5. Which do you go to? There's a Fibonacci-like rule.*

40.2 Quantum

Took a small pot of uranium from Amazon labelled 984 counts/minute. But that's an average,and quantum physics says we *can't* know. If Newton knew both where and velocity, he could predict where it would go. Bur $\Delta x \Delta p \geq \frac{h}{2}$.

But Einstein was not convinced about quantum physics.

40.3 Particles

First atoms, then protons etc., and now quarks. Is this the end, of is it "turtles all the way down".

We are losing information, and an astronomer far enough in the future would only see one galaxy. Similarly time before the big bang?

40.4 Consciousness

Can I know what your consciousness is like? "iPhone think therefore iPhone am" some time? There's a formula for the complexity needed for consciousness.

40.5 Mathematics

I took a Christmas cracker ($\nexists$ in USA). "What does the B in Benoit B Mandelbrot stand for". But Gödel's Incompleteness Theorem.

40.6 QA

Q Darwin: Our brains haven't evolved to know everything.

A True, brain capacity is finite. So are there proofs that will be too complicated to understand?

Q What do you do in mathematics?

A Group Theory and Number Theory. Group Theory is symmetry, but to understand primes we need a ζ function. I use ζ-functions to understand possible symmetries.

Chapter 41

NYU Data Science in Medicine/Health

9 June 2017.

41.1 NYU Center for Data Science: Claudio Silva

Idea to bring together all the pockets and small programmes: Courant, Tandon etc. Aim is to unify the educational programmes. Cross-school collaboration is not easy. Director of CDS reports direct to the Provost. We have several current Faculty and made some "core" hirings (these two total 17), and many (66) associated faculty. 2-year rigorous Master in Data Science commenced in Fall 2014, total 180 students: very close to 50:50 for gender. Most popular classes are ML, Big Data and Deep Learning. In Spring semester, about 50% of the students are not from DS programmes. In Fall 2017 we are starting a "track" system, e.g. DS+Physics. 297 applicants for 5 PhD places starting this year. 2 floors at 60 Fifth Avenue.

41.2 NYU Brain Initiative: Pesaran

Our research is to understand large circuits in the brain. Did his PhD on gravitational waves; way below sensitivity of measurements at the time. But they've now been found, so now believes that nothing is impossible to science.

The national Brain Initiative (Obama–2025) is a massive public-private partnership, Facebook, Google etc. have joined. Various startups in NeuroTech. One major area is "thinking about moving". Hence a range of Brain-Machine Interface challenges.

41.2.1 State of Art

2-photon calcium imaging can measure concentration of calcium in a cell, to $< 1\mu$ resolution. Shows a film of individual neurons firing. Commercial product.

Shows a reward/sensory/intent/command loop in the brain.

"Electronic skin" project: DARPA goal is (literally) millions of sensors, which he's doing with CMOS technology, and aims for device approval in three years. A human brain has 86G neurons and a mouse brain 8G.

Claims that translation to humans is merely one of scale.

41.3 Data Science at NYU Medicine

41.3.1 Where are we? (globally)

Image of Jabberwocky as AI, and hero as Medicine. This is an era of disruption.

41.3.2 NYU

Computational medicine; Bioinformatics; predictive analytics; population health; clinical departments; basic science departments; Tech4Health; Imaging.

41.3.3 Imaging

Biomedical imaging today is "art photography", taking shots. It's not a 21st century paradigm. Tomorrow is multimodal streaming data. Challenges ranging from segmentation and detection/classification to clinical workflow. Machines are pretty good at, for example, radiology detection. Of every 1000 women with breast scanning, 100 are recalled for further screening but only 5 have breast cancer.

Real-time analysis of image quality: should we let the patient go, or retake? Big challenge is designing next generation machines with AI in mind. Do we need images at all, or go direct from raw data to AI system??

41.3.4 New speaker

Central dogma is molecules + regulation. DNA/RNA are easy to measure, after that (proteins and changes to them) are harder to measure. We modify cells by adding proteins: fast (?but crude). Example of ERBB2 gene in breast/ovarian cancer. The behaviours seem very different. KRT5 again is different breast/ovarian/colon cancer. Correlate ERB2 with GRB7: high, but ERBB2 versus EBRR4 shows different behaviour.

41.4 Medical Image Reconstruction: Why should a Data Scientist care? Knoll

Prostate MRI is very difficult for multiparametric MRI, so this is one opportunity. But most current thinking starts with images. Can we use AI/ML to improve images from the raw data? Shows headlines from a lot of papers at a recent conference. Recent IEEE editorial on "Deep imaging" [JHD: nice buzzphrase]. Work on joint PET-MR imaging. MR acquires data in Fourier space.This process takes time, which impacts on patient experience (and data quality). Iterative: argmin .... Unfortunately my image models are too simple to capture complex anatomical details. What looks good to him is not good enough for a radiologist. So how do we separate artefacts from fine detail? Learn T gradient descent steps $u_t = u_{t-1} - \frac{\partial}{\partial u} \cdots$ (but discretised). Example of a very small feature on an image which is the real medical problem, as opposed to an artefact.

Acquiring data is a hard problem, but clinical images come with ground truth (JHD: really). Commercial data sets have been munged by proprietary systems: shows example that the commercial image does not map back to anything like the postulated raw data.

Real question is "what is in my data", not "what is the best image". Also, can we generalise to a larger patient cohort? Supported by MATLAB and CUDA code.

Q Artefact versus detail?

A Currently need radiologist input?

Q Same network or different one each time.

A Brain and knee seem to use the same, though the results aren't quite as good. The point is that it's the imaging hardware that generates the artefacts, and this doesn't change.

41.5 Machine Learning for Population Health:Narges Razavian

Partly ML, partly operations. Graph from Stanford showing Imagenet challenge, where between 2014 and 2015 AI started doing better than human (95%). NYU medical school: typical department has 1.5M unique patients, and 220M facts. 300K brain MRI across 90K patients, for example. Looking at some of our rare diseases, we detect other signals than had been previously detected. There are 250K+ NYUMC patients covered by 'phone calls and house visits to try to prevent kidney failure, heart attacks etc.

It's all very well for us to learn these models, but the doctors want to understand the biology. Answering this question involves understanding the ML

model. One exercise is to predict physician's behaviour, error rates. Also predict childhood obesity: environmental + EHR data.

Q ML for rare diseases?

A Good question. Sometimes, in some areas, we can pool data.

Q

A

41.6 Identifying Therapeutic Targets in Breast Cancer using Proteogenomics: Kelly Ruggles

Cancer patients acquire tumour-specific somatic variants. There are 825 Human Breast Tumours. [Ozenberger et al, Nature Genetics 45(2013) 1113-1120]. 77 of these we can do Proteomics/phosphoproteomics. Copy Number Alterations (CNA), Single Nucleotide Polymorphisms (SNPs), Novel Splice Junctions, Gene Expressions at the Genomics level. Three factors at Proteomics. Can proteogenomics guide discovery?

41.6.1 Proteomics

Discovery (measure all the protein expression, can enrich for phosphopeptides) and ??. NGS aids protein identification. "Black Sheep" is a project to identify aberrant proteogenomic events. Basically flagging outliers (even within the cancer data set). Found 181 phosphosite outlier kinases. Applied an outlier approach to drug study in patient derived Xenograft (PDX) tumours. Different tumours have different sensitivity to BKM120: why? Identified two kinases which prevented sensitivity to BKM120, which therefore show potential. Rapid. directly related to personalised drug treatment, and understandable by clinicians. Now looking an colon and ovarian as well.

Targeted proteomics: developed two tools to help out wet labs. Targeted MS cancerPanel with Washington U. 200kinases and 70 metabolic enzymes. Challenges: choose peptides to quantify each protein. We have developed PYTP picker for Proteotypic Peptide Selection. Also CRAFTS for Combinatorial Ratio Analysis For Targeted Spectrometry. Again in use in the wet labs. Currently can validly measure 173 kinases.

Lots of open questions. What is the best way to integrate the data and visualise the findings. What is the best investment for data collection. What about Metabolomics? Data sharing? Tool sharing?

Q Expand on visualisation questions? We've built awesome that aren't used any more.

A Off-line.

Q Unsupervised learning?

A A lot of what we've done is supervised. We've done some (not me personally).

41.7

41.7.1 multicompartment MR Fingerprinting via reweighted L_1 normalisation: Tang

The assumption is that only one fluid is present, but this is a problem clinically. If the assume that the mixture is sparse, then l_1 is a common technique.

41.7.2 Classification of Lung Cancer

200K cases/year in USA, half die in a year. How can we classify? 800K useful samples. 70% training, 15% validation and 15% testing. Used Google's CNN.

41.7.3 Understanding and Predicting Childhood Obesity

Standard methods are limited in predicting variance in childhood obesity models. Use 53K records from Lutheran Hospital Brooklyn. Future work to incorporate an ontology on diagnoses. RMS errors 2./47 (girls) 2.40 (boys). Two zip codes are very predictive:11210[1] and ?????.

41.7.4 NYU Data Catalogue

Not a repository. Some are external, e.g. census, and some generated by NYU. Important feature is "related data sets". This helps you find NYU experts on the datasets in question. Supports data sets split across multiple repositories. Working on adding software to the "other". Our metadata are interoperable with NIH.

[1] JHD postmeeting note "The people living in ZIP code 11210 are primarily black or African American. The number of people in their late 20s to early 40s is extremely large while the number of middle aged adults is large. There are also an extremely large number of single parents and a small number of families. The percentage of children under 18 living in the 11210 ZIP code is large compared to other areas of the country." from `https://www.unitedstateszipcodes.org/11210/`. That site gives the median household income as $55429 (no year), whereas 2014 figures show 55246 for New York State as a whole. However, it seems that `https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_1YR_S1902&prodType=table` shws (2015 figures) $78063 mean for Brooklyn, $78725 for Queens and $140997 for Manhattan. Note that this is mixing median and mean: those are all the data JHD can find.

41.7.5 Data Science at NWAS

41.7.6 Many–many relationships among urban spatial data

Started with NYC taxi trips data. Strong correlations with wind speed (actually massive dips in taxi correlated 100% with hurricane-sized spikes in wind speed.). But NYC data has over 300 spatial datasets, and one has > 300 columns.

41.7.7 Semantic segmentation of the primate brain

Currently need to label each voxel. Our system seems to perform really well. Next steps: combination; finer grain.

41.7.8 TIPseqHunter

A tool for detecting transposon intersections using TIPseq technology. Description of L1 junction repair, which JHD didn't follow. Used ML, describing validation.

41.8 Medical Image Analysis: from image data to quantitative information: Gerig

A 3D CT scan in $>$ 100MB. We expect healthcare data to grow 48%/year, and predicted to be 2.3 zettabytes by 2020. Databases and Data Federations. Hospitals, Picture Archiving Systems. Hospital Information Systems. There are ontology (JHD's word) and quality issues for images as well. Recommends Lego for calibration! With clinical data, anything that go wrong will! We have a non-sexy but useful data cleaning tool kit. Missing data is a problem, especially if its one slice in a 3D scan.

In these spaces we use Fréchet means. Also want longitudinal imaging: brain growth in Paediatrics, also aging and Neurodegeneration. Baseline versus follow-up etc. Hence what we need is spaciotemporal morphometry. We all understand regression, but how does one regress over images? Pick a specific age, say 40, then compute an average across the images at that age,then repeat. Also can look at individuals, e.g. Alzheimer's patients.

Claims, though, that regression is not applicable to longitudinal data. JHD didn't follow this argument. Looks at early brain development for autism research. Can see the growth in the visual cortex. A DL algorithm that uses surface area from MRI predicts diagnosis of autism (population was majority high-risk babies): 81% accuracy.

A study in Dynamic 3D Carpal Kinematics. Want to evaluate results after surgery.

So Statistics on Images+Patient Data = real clinical results (and papers!).

Q What is your big dream?

A Better standardisation of data.

41.9 Shalit

ML is great for detection, also prediction. Will this ICU patient die? But if we include this predictor in the workflow, then the workflow changes. See [CLG+15] — "Pneumonia patients with asthma are at less risk of dying" (because doctors automatically treat them more carefully!). Confuse "x predicts y" to "changing x to change y". Learning causality implies RCTs, which are expensive, difficult to control for minorities, unethical[2] and are not personalised. The alternative is observational studies. But this has, inter alia, the [CLG+15] effect. Needs a lot of de-confounding. For observational studies, we have to ask the counterfactual questions: what would have happened to patients who had the other treatments? Use DNN. Validating this is hard, but used a Labor Department data set.

Latent variables $z_t \in \mathbf{R}^d$, observations x_t and actions u_t. But NNs at every function. How do I validate a causal model? We can ask questions like "what happens to diabetics without their medication?" etc.

If your problem involves changing practice, you have a causal inference problem. This *might* by soluble this way.

Q How much data do you need?

A It depends. Validating personalised treatment is hard.

41.10 Panel Session

Moderator: Narges Razavian PhD in Math of image processing.

CS Tandon Ex EE, but ended up in Biomedical engineering, which involved hospital courses and experience. PH work, especially outside medicine: environment, diet etc.

Medicine EE/CS; Courant PhD.

Medicine EE. Director of Institute for Computational Medicine.

Hospital Also EE. Training programme: clinical data scientists? Research. Operational predictive data analytics team. Having a deployment perspective is important.

Consensus All of us had some kind of "embedding opportunities". How to systematise this?

Floor:CDS Masters students capstone projects?

Debate About proprietary nature of code and image data. Noted NIH tend to curate data, and will release unless the team is still actually working on it. One person said we always send things to bioarchive.

[2] As doctors etc. are forced to follow the RCT, rather than do "what is right".

But issues of privacy etc. over clinical data. One speaker noted the requirements for medical data, but said that anonymised data can be stored on a non-medical researchers system, and is being used for the HPC. But this won't work fro images attached to EHRs.

Also discussion on code sharing: getting others to use code is difficult.

Medic These days we don't have hypotheses, we just have data. But you need to know the biology.

? Google asked "can we automate the lab, even the discovery process"?

Chapter 42

On Voronoi Diagrams, Quadtrees and Lattices: Results in Geometric Algorithms: Huxley Bennett

PhD thesis defence 15 June, a student of Chee Yap. Good slide showing best papers in diagrams

42.1 Voronoi Diagrams via subdivision

Basic setting: n ppints in the plane: partition plane by closest point. Nearest neighbour search, motion planning etc. Can generalise from points, use different distances, different spaces.

Question 3 (What is "compute" a diagram?) *Geometric accuracy, topological accuracy, which model of computation (Real RAM is unrealistic).*

Example of three objects (triangle etc.), where changing the weights can change the topology of the VD.

Definition 9 *A minimization diagram partitions the plane into regions* $X_i = \{x \in \mathbf{R}^2 : f_i(x) \leq f_j(x) \forall j \neq i\}$

Hence weighted VD for bodies P_i is $f_i(x) = \inf_{p \in P_i} ||p - x||/w_i$.

So what we want is an *isotopic ϵ-approximation* $\tilde{M}(F)$ to $M(F)$. This was accepted by the conference, and deemed "reproducible" by the reproducibility referees.

1. Isolate Voronoi verties by root isolation

2. Connect vertices via Voronoi bisectiors by the marhcing algorithm.

Box predicates: "F has at least one root in B", ditto "most", also curvature predicates. Our implicit curves are $h_{i,j}^{-1}(0)$, where $h_{i,j} = f_i - f_j$. Use Poincaré–Miranda theorem to guarantee a common root of fmg is a box.

Example where the Voronoi regions aren't even connected.

42.2 Lattice Algorithms

Definition 10 *A lattice is the image span of linearly independent vectors.*

Question 4 *How similar are two lattices?*

Operator norm, condition number. Lattice distortion is the least condition number fo any mapping transform $\min_{T \in \mathbf{R}^{n \times n}} \{||T|||T(L_1) = T(L_2)\}$.

Example 31 (Rotation) *Distortion is 1, see [HR14: Regev]*

Example 32 *An example where $L{+}1 = L_2$ but has a bad basis, so the "obvious" map is badly conditioned.*

Problem 5 *Given two lattices and c, is there a mapping with condition number $\leq c$?*

We show this is NP-hard for approximations. Note "successive minima" λ_i.

Definition 11 $M(L_1, L_2) = \max_i \frac{\lambda_i(L2)}{\lambda}{}_i L_1)$.

Lemma 2 $D(L_1, L2) \geq M(L_1, L_2)M(L_2, L_1)$.

Intuitively, definte $T = B_2 B_1^{-1}$ for *reduced* bases, but traditional reduction isn't good enough: need Seysen-reduced bases. $S(B) = \max_i ||b_i|| \cdot ||b_i * ||$. [Sey93].

Chapter 43

Using Machine Learning to Study Neural Representations of Language Meaning: Mitchell

ACM Webinar, speaker from CMU. http://event.on24.com/eventRegistration/console/EventConsoleApollo.jsp?&eventid=1438516&sessionid=1&username=&partnerref=&format=fhaudio&mobile=false&flashsupportedmobiledevice=false&helpcenter=false&key=0E77DEF4D12BCB3FD968A0AD34317606&text_language_id=en&playerwidth=1000&playerheight=650&overwritelobby=y&eventuserid=171516914&contenttype=A&mediametricsessionid=138652983&mediametricid=2069613&usercd=171516914&mode=launch. See slides at [Mit17].

43.1 Word recognition

How does the brain recognise words, and how do they combine into sentences etc.? Shows images (four slices) of a brain hearing "bottle". Also mean activation, and then "bottle" less the mean activation. Can we train a machine-learning program to recognise these: tried SVM, Logistic, Deep net, Bayesian. Now that this works, we can ask whether classifiers work across people, which would imply similar decodes in the brain. Slide shows within and across participants, with very similar results, i.e. people decode the same way.

Then found English-Portuguese bilinguals, and found similar patterns. Also for words verses pictures. Easier to decode concrete nouns and emotions, rather than abstract nouns, or verbs (except when in context). Using a massive corpus,

measure correlations between verbs and nouns. Trained a model without 'celery' or 'airplane', then when asked to predict activity from co-occurrences did well (these two have very different co-occurrences. Give it a distinguisher test. On 9 subjects gets 0.79 accuracy, where 0.61 would have been 5% significant. Best predictor is a set of 20 features discovered themselves by ML, better than the 2018 from Mechanical Turk.

43.2 How long does it take

Answer is about 400ms, but fMRI isn't at this rate (roughly 1 second). MEG technology can do 1ms-resolution filming. Looked at the individual Mechanical Turk features over time. Word length appears in several regions of the brain at about 100ms, "graspability" at about 250ms.

43.3 Multiple words

Harry Potter stories, at a 500ms/word rate to separate them. Neural activity had a 75% chance of doing a discrimination between unseen sentences. [Webbe etal PLOS One 2014]. There's a chunk of the brain encoding the fact that dialogue is taking place.

Chapter 44

Finding Fibonacci: Devlin

Introduced as "The Math Guy from NPR".

> *The Quest to Rediscover the Forgotten Mathematical Genius Who Changed the World:* A Story about Books.

A Museum of Mathematics talk. Talk about a book, a book about a book, a book about a book about a book.

44.1 Books

Lessons learned included the value of physical artefacts: looking at 13th century books. Galileo might have held that book. We don't say "I wonder who used this keyboard last". Also the importance of books: Euclid, Galileo etc., largely as consolidators.

White writing "The Man of Numbers", I kept a log, and tracked my process of discovering how to do history. That's in "Finding Fibonacci". Noting also that what I learned about Fibonacci was parallel to what I knew about the history of Silicon Valley. That's in an e-book only: "Leonardo and Steve".

44.2 The Standard Story

Began in India, then Brahmagupta (7th century), then taken by Arab traders on the Silk Route back to North Africa and the Arab world, notably Baghdad (where modern algebra came from, invented to help traders, and make arithmetic more efficient), then Al-Khwarizmi and his book Al-Jabr. Then Leonardo of Pisa wrote [Leo02]: Pisa, Florence, Venice were the main centres of the Europe-rest of world trade. Leonardo's father was a major trader in Pisa. Moved his office to Bugia in modern Algeria.

"Trade was the killer app" for Hindu-Arabic arithmetic. Previously Roman Numerals, which were OK for addition/subtraction, but disastrous for multiplication/division. They had an abacus board. Also an elaborate system for finger

arithmetic, good up to 10,000. Note that neither method had any audit trail. These were pressing interests for Leonardo. We have two images of him, but not contemporary. The statue[1] is 1863, just an artist's conception. Father was called Guilichmus. Born around 1170, lived to c. 1250. Note that the 'bb' was deliberate, meaning it was going beyond the abacus. note that this wasn't the first introduction of Hindu-Arabic numbers as such, it was the rules for them.

> "Here beings the Book of Calculations composed by Leonardo Pisano Family Bonacci, in the year 1228". [KD's translation of preface of [Leo28]. Incipit abbaci Leonardi de domo filiou bonacii pisano ... a.m.cc.ii ... a.m.cc.xxviii. [JHD's reading of the manuscript]

The name was lost until Guillaume Libri [1838], who used Fibonacci to distinguish him from Da Vinci. Fibonacci sequence was so named by Lucas (because of the rabbit problem exercise). It was known to the Indians as well.

Also wrote a geometry textbook (which survives), a book on Number Theory (*Liber quadratorum* 1225) and a book of problems from his demonstrations at the Imperial Court. All these in Latin Also an abridged version of Liber Abbaci (not clear whether before or after second edition).

Note that there was an explosion of commerce: banks, insurance, trading empires, double-entry book-keeping(Medici) etc. immediately after the publications of Liber Abbaci. In the 1960s people started discovering a host of local (local dialect, lcoal weights and measures, local currency) abbacus books (250 studied). 1 pre-1300, 8 1300-1325, 10 1326-1350 etc.

Note that, if you wanted a book, you would make a copy (in haste, slavishly), then you'd annotate to as you study, then maybe rewrite, and this would then have descendants etc. Earliest known Abbacus school was 1294, and at one point there were 20 in Florence. These are all very similar to each other, but not to Liber Abbaci. These books also have some material from *Practica geometriae*. Hence, if there were an independent author who started the pamphlet-style books

Goetzmann (Yale) traces all modern finance back to Liber Abbaci.

44.2.1 Discovery 2003

Rafaella Franci found, in Via dè Ginari 10, Florence, a copy (1290) of a book by an unknown Umbrian. The person who wrote it clearly didn't understand it completely, at least at first copying. "This is the book of abbacus, according to the master Leonardo of Pisa".

There has been another manuscript discovered. This has a pigeon sequence problem instead. The Pisan mansuscipt is better organised than the Umbrian.

[1] Piazza dei Miracoli.

44.3 Why was he forgottem?

Victim of technology. One of first books printed was a book of arithmetic, in Treviso. The printer took a most recent abbacus book. Once we had printing, no-one would go back to the origins. Luca Pacioli's *Summa de arithmetica* "And since we follow for the most part Leonardo of Pisa". Pietro Cossali saw this, c. 1800, and that started the history chase.

"I have spent much of my career writing accessible books, hence my admiration for Fibonacci."

44.4 Explore Pisa

And try to put myself in his shoes? Pictures of Pisa.

Port of Livorno. The medieval Porto Pisano.

The statue was preserved despite being in the centre of the battle over the Arno in the second world war. Finally discovered statue in CampoSanto.

Picture of 19th century plaque commemorating the citizenship of Pisa (decree dated 1241, so presumably Fibonacci was still alive then), kept in Palazzo Toscanelli, which is where Byron stayed.

The Italian Computer Society is at 12 via Fibonacci in Florence.

Siena manuscript has drawings of the finger arithmetic system he is replacing. There were no symbols, partly because words are error-correcting (think uncomprehending copyists), and symbols aren't. Symbols were sometimes put in the margins, but only as glosses. Note that printing essentially reversed this. Boncompagni has a printed version (18??) of the Florence manuscript. Laurence Sigler produced an English translation (died before finished, disc recovery issues etc.).

I noted that Jobs recognised the potential of the [PARC] system just as Fibonacci recognised the potential of Hindu-Arabic arithmetic. Jobs first produced a clunky machine, then a simpler one, just as Fibonacci did!

44.5 Q&A

Q Cover on Man of Numbers? starts 1,1,2,3, then goes wrong.

A Publisher's visual pattern.

Q You handled the original?

A No, an early copy. Siena is probably the oldest, 13th century. It took me several attempts to read it, as they'd essentially lost it. Not allowed to photograph it, but they would take images for me. Florence was rather harder. Stanford Faculty card got me into see the entrance of the archives. There was a Yorkshire assistant there who got him a National Archives Access Card (lifetime pass). History is a very human business!

Q Who funded him? Medici? Why Jobs: what about Jeff Raskin.

A Jobs is the figurehead of popular history.

***** I could write this because Franci gave me her manuscript, which for her was describing an interesting, if minor, piece of history.

Chapter 45

Conference in honour of Gérard Ben Arous

JHD attended a few talks.

45.1 Heat Kernel Estimates for Liouville Brownian Motion: Ofer Zeitoumi

Remininiscences of Gérard at another birthday do.

Brownian motion W_t in $\mathbf{R}^d$, gnerator $\Delta/2$: $p_t(x,y) = \frac{1}{\sqrt{2\pi t}^d} exp^{-|x-y|^2/2t}$.

Theorem 22 (Varadhan 1975) *With uniform elliptic generator, the heat kernel has*

$$t \log p_t(x,y) \to_{t=0} \dots$$

Focus on 2D torus. Intrest in cases when $V(\cdot)$ is not smooth, maybenot even pointwise defined. Defined as distribution, the *Gaussian free field*. Then the measures $\mu_\epsilon^\gamma(dx)$ converge, if $\gamma < 2$, to a formally defined limit μ_V. This is supported on γ-thick points $\{x : V_\epsilon(x)/\log(1/\epsilon) \to_{\epsilon\to 0} \cdots$.

Given a μ_V, we can look at the geometry associated with it. A general paradigm, verified for BM on many fractals, is tha the heat kernel should behave, for short time, , depedning on d_H the Hausdorff dimension and $2d_H/b$ teh spectral dimension. To have any hope of identifying distances, we need to find d_H.

Theorem 23 (Watabiki 1993)

$$d_H = 1 + \frac{\gamma^2}{4} + \sqrt{\left(1 + \frac{\gamma^2}{4}\right)^2 + \gamma^2}$$

45.2 Some demonstrations of universality: Percy Deift and Tom Trogdon

See [DMOT14, DT17] Solutions of completely integrable Hamiltonian systems appear everywhere when we consider random matrices. QR etc. for eigenvalues are also really completely integable Hamiltonian systems. What happens if we merge these two facts? What happens when we compute the eigenvalues of a random matrix?

Consider a matrix M in block form, where the first element is a $k \times k$. If in this format the off-diagonals are $\leq \epsilon$, then the eigenvalues of the two diagonal matrices are, within ϵ, the eigenvalues of the whole matrix. This process is deflation, and deflation time is the time taken to achieve it. Compares two algorithms, QR and Toda. Two ensembles of random matrices: BE of iid zero-mean Bernouilli random variables, and GOE — iid mean zero normals, $\epsilon = 10^{-10}$ and $N = 100$. The ensembles behave the same, but the graphs for the two algorithms are different.

Start looking at a variety of finite dimensional numerical problems. We find universality. Then what happens in infinite dimensional processes? But these are deterministic processes with random input. Then we looked at a genetic algorithm.

Always look at distribution of $(x_i - \mu)/\sigma$. Need N large and ϵ small. Same curve with purely random matrices and those with geometric structure. Then looks at genetic algorithms. There are choices of mutation/cross-over ratios etc., but this seems to lead to the same universal graphs.

Human experiment. 45 participants shown 200 pictures, pairs of images (each say, 9 dots) to decide which is bigger. What's actually recorded is time-to-decision. All look like the shifted/scaled Gumbel distribution.

Also went to Google looking at search time, for English and for Turkish words $f(x) = \sigma g(\sigma x + \mu)$ where $g(x) = \exp(-x - e^{-x})$.

45.2.1 So much for experiments

What's the theory? There's a distribution of the inverse gap between the largest eigenvalue and the second largest.

Claims that this is related to the natural process whereby we automatically adjust for scale when comparing objects.

Data for `mortality.org`, where all countries examined (SWE, IRL, US, CHL) were the same, but Russia different: which he blamed on the male vodka problem.

Chapter 46

From Hopf Algebras to Machine Learning via Rough Paths: Lyons

Share a passion for applications as well as theory with Gérard. This is basically a talk aout streamed data. In the Mathematics of Information, sometimes order of data matters (book, tick data, astronomical data) and sometimes doesn't (census data, cross-sectional data). The second os largely topological, and not my concern today.

46.1 Theory

Newton said we should look at where a path is going, so we can regard a path as a partition and chords. Note that the chords are an Abelian summary (order doesn't matter) in terms of the destination. We tried to look at the solar system this way, but there are an unbounded number of planetoids etc., generating a high-frequency fluctuations. Hence stochastic Calculs and Itô. $y_{i=1} - i_i = f(y_i, x_i)(x_{i+1} - x_i)$. this is a semi-martingale almost surely.

Photograph of Merton/Scholes at ICMS993, before Nobel prizes. If an obejct is not tradeable, then semi-martingales do not tell the story.

Most learning is based on linear regression, and MacLaurin is basically saying that we want is a *linear* combination of basic functions. What abut functions on streamed data. Each datum has an effect, and these compose, so data are a group. Then every path has a representation as a grouplike element. Unlike Neural Nets, this sis a graded algebra. Claims that this is faithful for bounded paths and rough paths, under a suitable definition of equivalence. So we have a transform path $\to$ signature information.

46.2 Real Applications

Team involving ATI. Three areas:

- Example NMIS data, turned from pictures to streamed information. Compute a number of terms in the signature, and give them to a linear NN classifier (which shouldn't care about having a linear transform, but does). Also action classification (a standard ML example data set). Use a stick man approach: paths in 24D. Facila Expression Recognition on CK+ data: again evolution of paths taken by landmarks.

- handwriting recognition: Chinese handwriting finger on screen. Colour each point with the shape of the local path. Done with the Chinese 1–5M downloads of Android version. Oddly: our Chinese collaborators have no incentive to improve, as we don't want to encourage people to write badly!

- evolution of mental health. Data from a clinical trail where patients created a path. Can we tell healthy/bipolar disorder/personality disorder diagnosis (by CT experts) at the start of the treatment. Path was a score from 0–6 in 7D (emotions w.r.t. mobile 'phones). Split into 20-day chunks. Then various "tricks", a random forest classifier, and a predictor machine. Agreed on all but one person (under investigation).

Chapter 47

Intelligent Question Answering Using the Wisdom of the Crowd: Preslav Nakov

47.1 Abstract

In recent years, community Question Answering forums such as StackOverflow, Quora, Qatar Living, etc. have gained a lot of popularity. As such forums are typically not moderated, this results in noisy and redundant content; yet, they are highly valued by users as a source of information. I will explore three general problems related to such forums, focusing on Qatar Living: (i) deciding which answers are good, (ii) finding related/duplicated questions, and (iii) finding good answers to a new question. This will involve models based on deep learning and semantic/syntactic kernels. I will further discuss extensions of this work in directions such as application to Arabic, cross-language question answering, fact checking, trollness detection, answer justification, and interactive question answering.

Bio: Dr. Preslav Nakov is a Senior Scientist at the Qatar Computing Research Institute, HBKU. His research interests include computational linguistics, machine translation, question answering, sentiment analysis, lexical semantics, Web as a corpus, and biomedical text processing. Preslav Nakov received a PhD degree in Computer Science from the University of California at Berkeley. He is Secretary of ACL SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics. He is also a Member of the Editorial Board of the Journal of Natural Language Engineering and an Associate Editor of the AI Communications journal. Preslav Nakov co-authored a Morgan & Claypool book on Semantic Relations between Nominals, two books

on computer algorithms, and many research papers in top-tier conferences and journals.

47.2 Talk

"Why do we need question-answering if we have Google". "How old is Donald Trump" does well, but "Can somebody tell me how old Donald Trump is?" doesn't get good answers, also example about "Qatari residence permits". 8 people at QSRI working with five at MIT. Our goal is to suppose complex questions, and interactions. Neither Google nor Siri can do these. Starting point was an "intelligent Qatar Tourist Browser", asking questions like "Cheap Italian Restaurants in Doha". Need seminatic parsing. Start with semi-Markov CRF, re-rank with kernels. Note that semantics parsing is at the heart of chatbots. Standard design is input $\rightarrow_{NLP}$ structured query $\rightarrow$ database $\rightarrow$ text output.

For question-answering, we want immediate answers if they exist, also removal of duplicates etc. Real example: multiple questions in one "query", in broken English. Two good answers, two bad ones, and two collateral remarks in the database. So answer ranking (A), question similarity (B) and Answer Selection (C).

A use a pairwise network for comparison. Note that early questions in threads tend to be better, Features specific to questions, answers, the pair, and metadata (X gives good answers, etc.). We have hand-labelled data. Example: a "thank you" from the questioner is valuable information. See papers in RANLP '17.

Also fact-checking

1. Prioritisation
2. general claims
3. ...

A good example of contrary answers is medical/diet interaction. Look at Cochrane. Again a set of neura; net categorisers.

Also Q/A in Arabic based on same English-language database. Solution based on tree kernels. This requires cross-language question-question comparison.

Quotes Yann Le Cun on "Generative Adversarial Nets".

Returning to chatbots, we extracting Q/A sentence pairs, and trained a seq2seq model, optimising it for the chatbot QA scenario: See also Microsoft's QnA Maker. That's a FAQ→chatbot constructor. `http://130/204/203/.149:5000/static/index.html` — showed some humorous examples: reminded JHD very much of Parry/Doctor [Cer73].

Q This isn't really "wisdom of the crowd", which is averaging the answer.

A We do some of that.

A Note that we aren't really doing deep factuality checking. There's also a temprality question: visa rules change over time.

47.2.1

47.2.2

Bibliography

[Ama17] Amazon Web Services Inc. Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region. `https://aws.amazon.com/message/41926/`, 2017.

[Ano16] Anonymous. Ethereum/TheDAO hack simplified. `http://blog.erratasec.com/2016/06/etheriumdao-hack-similfied.Ethereum/TheDAO%20hack%20simplified`, 2016.

[BKM08] Peter A. Boncz, Martin L. Kersten, and Stefan Manegold. Breaking the Memory Wall in MonetDB. *Commun. ACM*, 51(12):77–85, December 2008.

[BSV17] E. Balkanski, U. Syed, and S. Vassilvitskii. Statistical Cost Sharing. `https://arxiv.org/abs/1703.03111`, 2017.

[But16] V. Buterin. CRITICAL UPDATE Re: DAO Vulnerability. `https://blog.ethereum.org/2016/06/17/critical-update-re-dao-vulnerability/`, 2016.

[Cer73] V.G. Cerf. PARRY Encounters the DOCTOR. *The Internet Activities Board*, 1973.

[Chi09] A.L. Chistov. Double-exponential lower bound for the degree of any system of generators of a polynomial prime ideal. *St. Petersburg Math. J.*, 20:983–1001, 2009.

[CLG+15] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM*, pages 1721–1730, 2015.

[DMOT14] P. Deift, G. Menon, S. Olver, and T. Trogdon. Universality in numerical computations with random data. *Proc. Natl. Acad. Sci. USA,*, 111:14973–14978, 2014.

[DT17] P. Deift and T. Trogdon. Universality in numerical computation with random data. Case studies, analytic results and some speculations. `https://arxiv.org/abs/1703.08092`, 2017.

[ES13] I. Eyal and E.G. Sirer. Majority is not enough: Bitcoin mining is vulnerable. http://arxiv.org/abs/1311.0243, 2013.

[Fen13] R. Feng. Notes on Hrushovski's Algorithm for Computing the Galois Group of a Linear Differential Equation. http://arxiv.org/abs/1312.5029, 2013.

[Fra95] P. Frankl. Extremal set systems. *Chapter 24 in The Handbook of Combinatorics*, 1995.

[GGOW15] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. A deterministic polynomial time algorithm for non-commutative rational identity testing. http://arxiv.org/abs/1511.03730, 2015.

[GGOW16] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. A Deterministic Polynomial Time Algorithm for Non-commutative Rational Identity Testing. *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 109–117, 2016.

[Gil59] D.B. Gillies. Solutions to general non-zero-sum games. *Contributions to the Theory of Games 4o*, 4:47–85, 1959.

[Her26] G. Hermann. Die Frage der Endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[KA16] G. Karame and E. Androulaki. Bitcoin and blockchain security. *Artech House*, 2016.

[KLR16] A. Klawonn, M. Lanser, and O. Rheinbach. A nonlinear FETI-DP method with an inexact coarse problem. *In Domain Decomposition Methods in Science and Engineering XXII*, pages 41–52, 2016.

[Leo02] Leonardo di Pisa (Fibonacci). Liber Abbaci. *Manuscript*, 1202.

[Leo28] Leonardo di Pisa (Fibonacci). Liber Abbaci (second edition). *Manuscript*, 1228.

[LS16] S. Li and T. Shemyakova, E. andVoronov. Differential operators on the superline, Berezinians, and Darboux transformations. https://arxiv.org/abs/1605.07286, 2016.

[McG11] R. McGregor. Zhou's cryptic caution lost in translation. https://www.ft.com/content/74916db6-938d-11e0-922e-00144feab49a, 2011.

[ME17] M. Mueller-Eberstein. The Next Radical Internet Transformation: How Blockchain Technology is transforming Business, Governments, Computing and Security models. https://on.acm.org/t/the-next-radical-internet-transformation-how-blockchain-technology-is-transformi
51, 2017.

[Mit17] K. Mitchell. Using Machine Learning to Study Neural Representations of Language Meaning (slides). `http://event.lvl3.on24.com//event/14/38/51/6/rt/1/documents/resourceList1497536590366/resourcesmitchell_acmwebinar_june2017.pdf`, 2017.

[NJ12] K. Nichols and V. Jacobson. Controlling Queue Delay. *Comm. ACM 7*, 55:42–50, 2012.

[Pag17] C. Page. Major outage on AWS S3 causes havoc for millions. `http://www.computing.co.uk/ctg/news/3005594/major-outage-on-aws-s3-causes-havoc-for-millions`, 2017.

[PD15] M.J. Paul and M. Dredze. SPRITE: Generalizing Topic Models with Structured Priors. *Transactions of the Association for Computational Linguistics*, 3:43–57, 2015.

[Reu96] C. Reutenauer. Inversion height in free fields. *Selecta Mathematica New Series*, 2:93–109, 1996.

[SCB+16] A. Sivaraman, A. Cheung, M. Budiu, C. Kim, M. Alizadeh, H. Balakrishnan, G. Varghese, N. McKeown, and S. Licking. Packet transactions: High-level programming for line-rate switches. *In Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference*, pages 15–28, 2016.

[Sey93] Martin Seysen. Simultaneous reduction of a lattice basis and its reciprocal basis. *Combinatorica*, 13(3):363–376, 1993.

[SSA+16] A. Sivaraman, S. Subramanian, M. Alizadeh, S. Chole, S.T. Chuang, A. Agrawal, H. Balakrishnan, T. Edsall, S. Katti, and N. McKeown. Programmable Packet Scheduling at Line Rate. *In Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference*, pages 44–57, 2016.

[YTCK16] O. Yair, R. Talmon, R.R. Coifman, and I.G. Kevrekidis. No equations, no parameters, no variables: data, and the reconstruction of normal forms by learning informed observation geometries. `https://arxiv.org/abs/1612.03195`, 2016.