Scattered thoughts from applied probability

David Aldous

14 June 2019

Topic 1: Should you do the back-of-an-envelope calculation before the multi-million dollar project?

Non-transitive dice are a "paradox" in the sense that one might just assume such things are impossible without thinking about it. I'll talk about another such paradox.

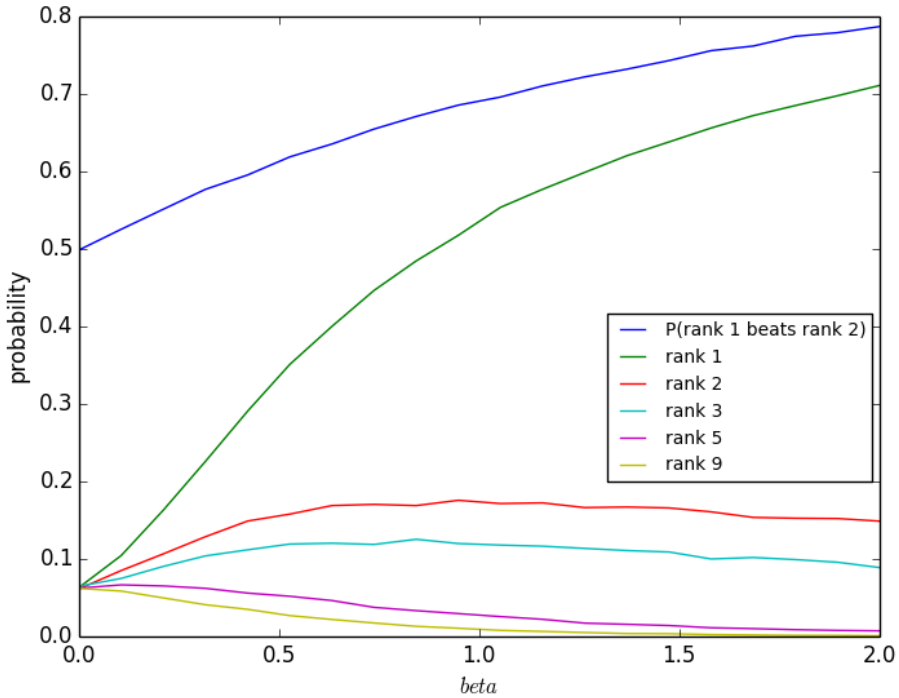
Background analogy: in a sports match the better team doesn't always win, but is likely to win. So in a sports tournament the probabilities of different final winners should be ordered as the abilities of the different teams.

Let's briefly say a math model to check this (details not important). In the Bradley-Terry model

$$\mathbb{P}(A \text{ beats } B) = F(x_A - x_B), \quad x = \text{ ability, } F = \text{ logistic.}$$

Make a probability model of random abilities, with a parameter controlling variability of abilities, and simulate a 16 team single-elimination tournament.

Figure: Probabilities of different-ranked players winning the tournament, compared with probability that rank-1 player beats rank-2 player (top curve).



Here math is consistent with common sense.

In a **prediction tournament** contestants state probabilities of future geopolitical events. Here are 5 out of 80 questions asked currently on `gjopen.com`.

• Will an armed group from South Sudan engage in a campaign that systematically kills 1,000 or more civilians during 2019?
• Will there be a lethal confrontation in the South or East China Sea between the military forces, militia, or law enforcement personnel of China and another country before 1 January 2020?
• Before 1 October 2019, will the U.S. House of Representatives pass an article of impeachment against President Trump?
• Will North Korea launch an intercontinental ballistic missile (ICBM) before 1 January 2020?
• Between 22 February and 31 December 2019, will more than one CRISPR gene-edited baby be born?
• Before 31 December 2019, will Fitch, Moody's, or S&P downgrade the United Kingdom's long-term local or foreign currency issuer ratings?

DARPA has a shyer cousin IARPA – non-classified research of indirect interest to the Intelligence community. They funded a series of **Good Judgment Projects** in which volunteers (including me) as individuals and teams make forecasts for such questions.

The point is to gather evidence and expert opinions before giving an answer – and (unlike an exam) there are no limitations – you can copy other people's answers, or if you happen to be a personal friend of Vladimir Putin . . . . . .

Important: contestants are not asked to give a Yes/No prediction, but instead are asked to give a numerical probability, and to update as time passes and relevant news/analysis appears.

Call for one 2018 contest

IARPA is looking for approaches from non-traditional sources that would improve the accuracy and timeliness of geopolitical forecasts. IARPA hosts these challenges in order to identify ways that individuals, academia, and others with a passion for forecasting can showcase their skills easily.

Why Should You Participate: This challenge gives you a chance to join a community of leading experts to advance your research, contribute to global security and humanitarian activities, and compete for cash prizes. This is your chance to test your forecasting skills and prove yourself against the state-of-the-art, and to demonstrate your superiority over political pundits. By participating, you may:

- Network with collaborators and experts to advance your research
- Gain recognition for your work and your methods
- Test your method against state-of-the-art methods
- Win prizes from a total prize purse of $200,000

Why are millions of taxpayer dollars being spent running such projects?

- What makes some individuals better than others? The study starts with a lengthy test of "cognitive style" to see what correlates.
- What makes some teams better than others? How to combine different sources of uncertain information/analysis is a major issue Intelligence assessment. The project managers see team discussions.

How can we assess someone's ability? We do what Carl Friedrich Gauss said 200 years ago – use **mean square error** MSE. An event is a 0 - 1 variable; if we predict 70% probability then our "squared error" is
(if event happens) $(1.0 - 0.70)^2 = 0.09$
(if event doesn't happen) $(0.0 - 0.70)^2 = 0.49$

As in golf, you are trying to get a low score. A prediction tournament is like a golf tournament where no-one knows "par". That is, you can assess people's relative abilities, but you cannot assess absolute abilities.

Writing S for your "tournament score" when the true probabilities of the n events are $(p_i, 1 \le i \le n)$ and you predict $(q_i, 1 \le i \le n)$,

$$\mathbb{E}S = \sum_i p_i(1 - p_i) + n\sigma^2 \tag{1}$$

where

$$\sigma^2 := n^{-1} \sum_i (q_i - p_i)^2$$

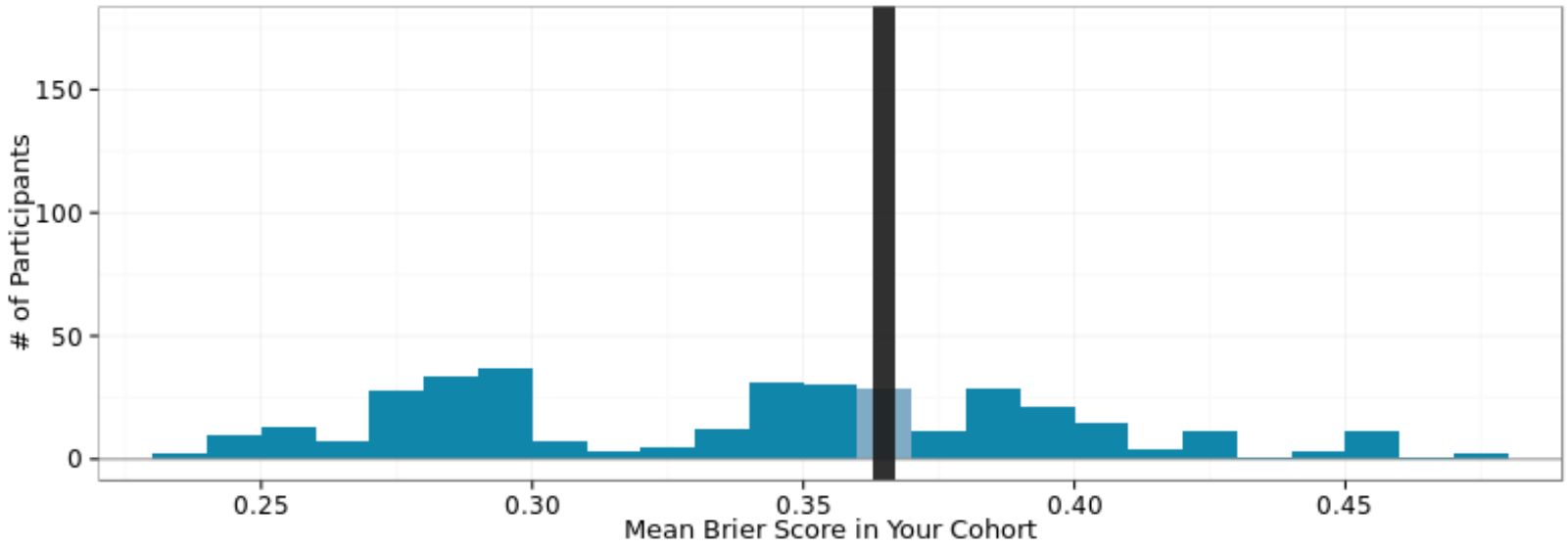
is your MSE (mean squared error) in assessing the probabilities.
So for contestants A and B

$$n^{-1}\mathbb{E}(S_A - S_B) = \sigma_A^2 - \sigma_B^2$$

and so in the long run we can tell who is the more accurate forecaster.

This has philosophical interest, best discussed over beer.

Here is a histogram of 2×scores of individuals in the 2013-14 season `GJP challenge`. The season scores were based on 144 questions, and a back-of-an-envelope calculation gives the MSE due to intrinsic randomness of outcomes as around 0.02, which is much smaller than the spread observed in the histogram. The key conclusion is that there is wide variability between players – as in golf, some people are just much better than others at forecasting these geopolitical events.



In the long run we could tell who is the more accurate forecaster, but what about chance variation in realistic-size tournaments? We need a **model** for comparing contestants scores.

- 100 questions
- true probabilities (unknown to contestants) uniformly spread from 5% to 95%.
- For each contestant A there is a RMS error σ_A for their predicted probabilities: that is, in the model, for each event the prediction $p_{predicted}$ by A is random and such that

$$\sigma_A^2 = \mathbb{E}(p_{predicted} - p_{true})^2.$$

- (complete model specification discussed later)

Now we can simulate the tournament.

Figure: One-on-one comparison: Chance of more accurate forecaster beating less accurate forecaster in 100-question tournament.

		RMS error (less accurate)					
		0.05	0.1	0.15	0.2	0.25	0.3
	0	0.73	0.87	0.95	0.99	1.00	1.00
RMS	0.05		0.77	0.92	0.97	0.99	1.00
error	0.1			0.78	0.92	0.97	0.99
(more	0.15				0.76	0.92	0.97
(accurate)	0.2					0.76	0.91
	0.25						0.73

So this is quite similar to Bradley-Terry: use the RMS probability-prediction error as "ability", and roughly

$$\mathbb{P}(A \text{ beats } B \text{ in prediction tournament }) \approx F(\sigma_A - \sigma_B)$$

for some function F

A leader in this field is Philip Tetlock, with a popular book *Superforecasting* and a 2017 *Science* article and a 2015 paper *Identifying and cultivating superforecasters as a method of improving probabilistic predictions*. They write

> *[the winning strategy for teams over several successive tournaments was] culling off top performers each year and assigning them into elite teams of superforecasters. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics.*

Of course this is essentially the same way that professional football players – or mathematics professors – are developed.

But let's check this holds up mathematically in our prediction context.

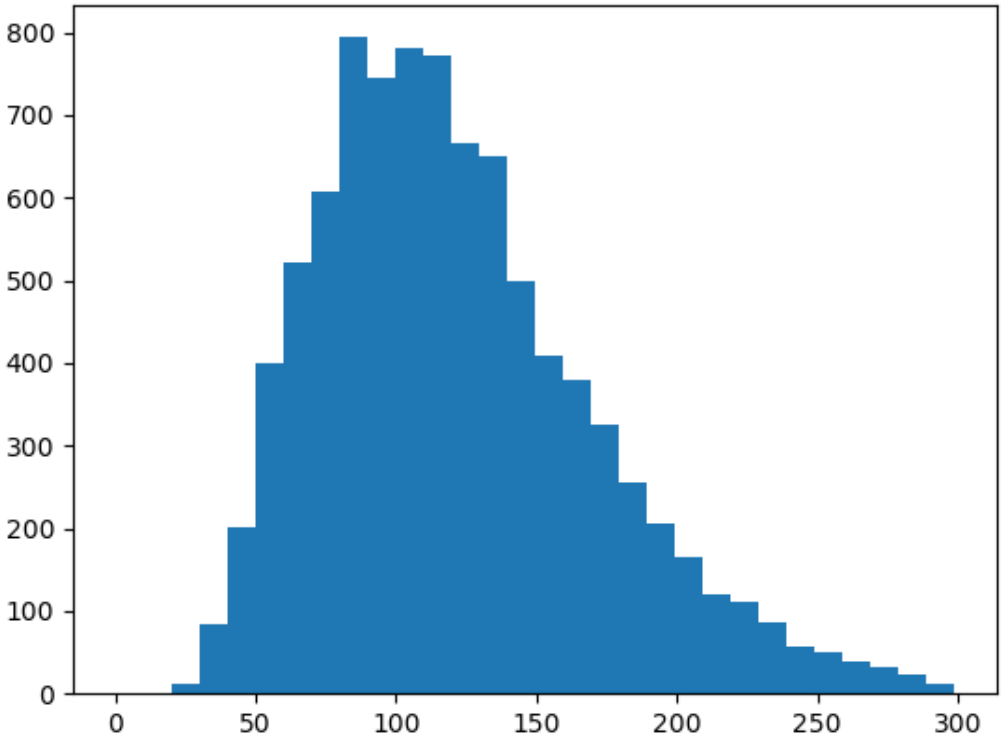
Recall "ability" of contestant formalized as RMS error σ in predicting probability. For a **tournament model** we need a model for variability of ability over contestants:

- 300 contestants
- σ varies evenly from 0 to 0.3.

So we can rank contestants from 1 to 300 in terms of ability. For a tournament with a million events, by LLN the order of scores would closely match the ranking of ability. But what about a realistic size tournament with 100 events?

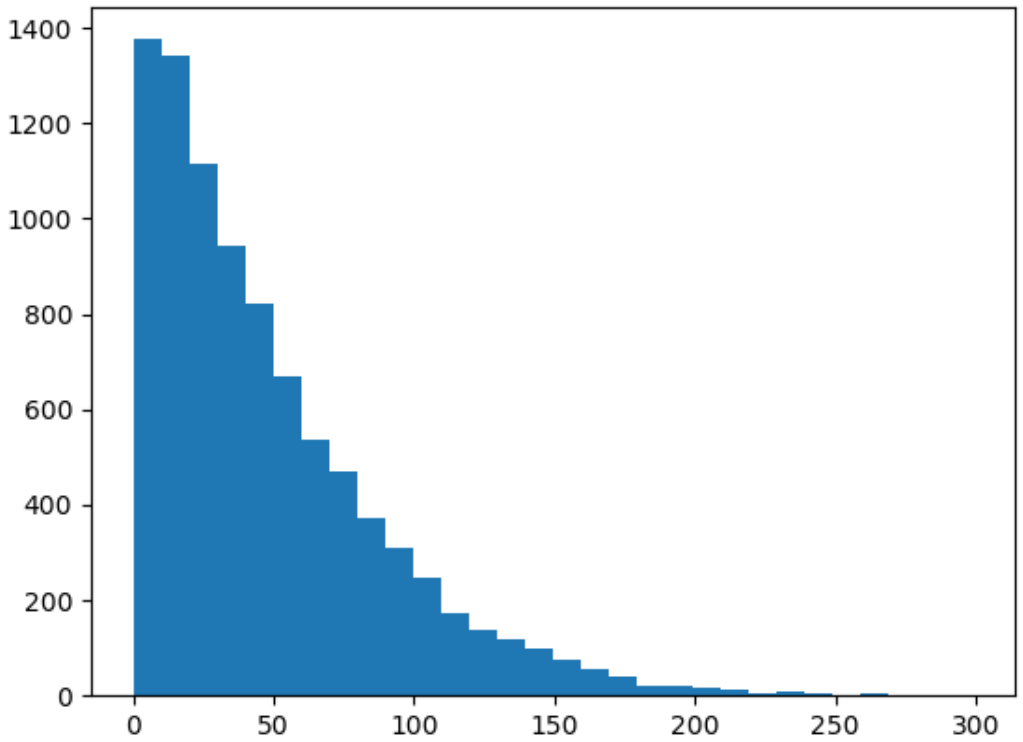
Specifically, what is the (ability) rank of the tournament winner?

Here is the first simulation I did.



Maybe something wrong with my amateur Python code?

Maybe no-one is near-perfect in predicting probabilities? Here are results if the abilities (RMS errors) σ range over $[0.1, 0.4]$ instead of $[0, 0.3]$



This is partly in line with common sense – the best forecasters are relatively more likely to win – but still the winner is liable to be around the 50th best contestant.

So that's the paradox – according to this model, tournaments are a surprisingly ineffective way of identifying the best forecasters, even though IARPA is spending millions of dollars doing precisely this.

Now the issues are

- Is there a calculation to qualitatively explain these simulation model results?
- Why are our model results very different from what is claimed for real tournament results?

The back-of-envelope calculation

Consider a 100-question tournament in which the true probabilities are all 0.5. What are the scores S?

- A perfectly accurate forecaster: $S = 25.0$.
- A contestant who predicts 0.4 or 0.6 randomly on each question: $\mathbb{E}S = 26.0$, s.d.$(S) = 0.98$.
- A contestant who predicts 0.3 or 0.7 randomly on each question: $\mathbb{E}S = 29.0$, s.d.$(S) = 1.83$.

Moreover, as a special feature of the "all true probabilities are 0.5" setting, different contestants' scores are independent. In our simulated setting of 300 contestants, some scores will by chance be around 3 s.d.'s below expectation. With RMS prediction errors ranging from 0 to 0.3, we expect a winning score around 23 and we will not be surprised if this comes from the 100th or 200th best forecaster.

Why is this happening? The key point is that for predicting probabilities the expected cost of small errors scales as $(\text{error})^2$ while the s.d. scales as (error). This is quite different from a typical sport – golf or basketball – where the winner is decided by point difference, points earned in some success/failure way. In sports the expected point difference scales as (difference in ability) and the s.d. of score is roughly constant.

A superficial conclusion of our results is that winning a prediction tournament is strong evidence of superior ability *only* when the better forecasters' predictions are *not* reliably close to the true probabilities. But are our models realistic enough to be meaningful? Two features of our model are unrealistic. One is that contestants have no systematic bias towards too-high or too-low forecasts. But alternate models allowing that give roughly similar results.

I guess the most serious issue is that the errors are assumed independent over both questions and contestants. In reality, if all contestants are making judgments on the same evidence, then (to the extent that relevant evidence is incompletely known) there is surely a tendency for most contestants to be biased in the same direction on any given question. Implicit in our model is that, in a large tournament, this "independence of errors" assumption means that different contestants will explore somewhat uniformly over the space of possible prediction sequences close to the true probabilities, whereas in reality one imagines the deviations would be highly non-uniform.

Statistical analysis of real tournament data is too complicated (for me). But here are 2 data points.

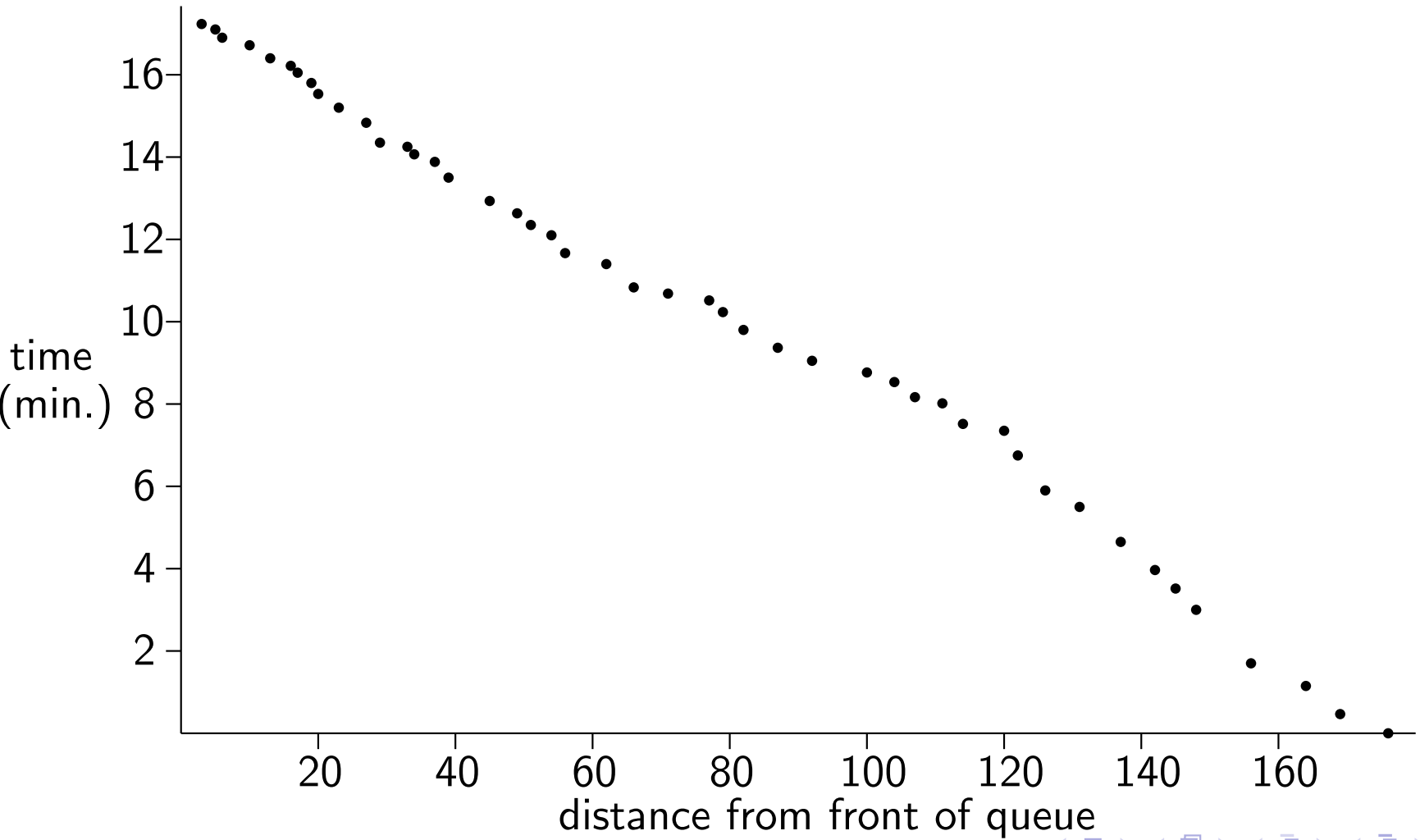
[start next simulation]

I tell students: keep your eyes open for phenomena that you might model stochastically.

Here's an example that led to a plausible explanation of observations, and some cute math.

Topic 2: Stop-and-Go at Airport Security

17 minutes in line at security at Oakland airport



This phenomenon is easy to understand qualitatively. When a person leaves the checkpoint, the next person moves up to the checkpoint, the next person moves up and stops behind the now-first person, and so on, but this "wave" of motion often does not extend through the entire long line; instead, some person will move only a short distance, and the person behind will decide not to move at all.

Intuitively, when you are around the k'th position in line, there must be some number $a(k)$

- $a(k) =$ average time between your moves
- $a(k) =$ average distance you move when you do move
- $\mathbb{P}(W > k) = 1/a(k)$ for length of typical wave.

You are moving forwards at average speed 1 [unit time = service time, unit distance = average distance between people in queue]. This immediately suggests the question of how fast $a(k)$ grows with k.

I will present a stochastic model in which $a(k)$ grows as order $k^{1/2}$.

In classical *queueing theory* randomness enters via assumed randomness of arrival and service times. In contrast, even though we are modeling a literal queue, randomness in our model arises in a quite different way, via each customer's choice of exactly how far behind the preceding customer they choose to stand, after each move. That is, we assume that "how far behind" is chosen (independently for each person and time) from a given probability density function μ on an interval $[c_-, c^+]$ where $c_- > 0$. We interpret this interval as a "comfort zone" for proximity to other people. By scaling we may assume μ has mean 1, and then (excluding the deterministic case) μ has some variance $0 < \sigma^2 < \infty$.
In words, the model is

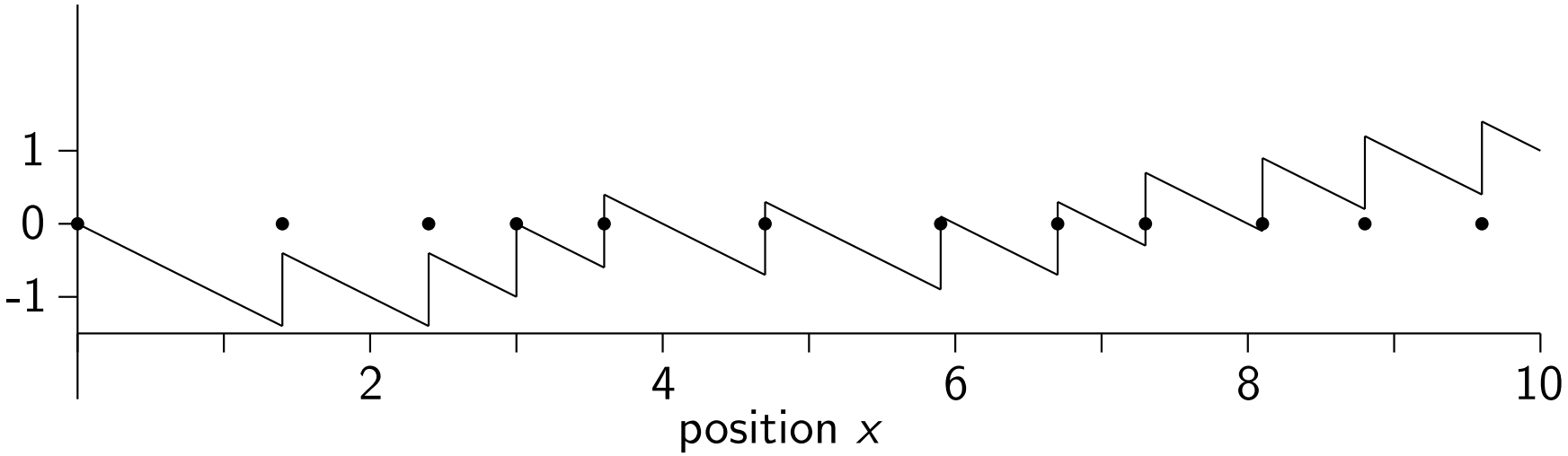
> *when the person in front of you moves forward to a new position, then you move to a new position at a random distance (chosen from distribution μ) behind them, unless their new position is less than distance c^+ in front of your existing position, in which case you don't move, and therefore nobody behind you moves.*

- Model **could** have been studied 60 years ago – but I can't find any closely related literature. Some traffic models loosely similar; also TASEP.
- Model as infinite queue.
- You might guess process has stationary distribution with inter-customer distances IID μ – **no**.
- Not obvious how to start analysis.
- Seem obvious that process time-converges to some unique stationary distribution – **cannot prove**.

It turns out there is a non-obvious picture which explains (intuitively) the $k^{1/2}$ scaling in this model.

A configuration $\mathbf{x} = (0 = x_0 < x_1 < x_2 < x_3 \ldots)$ of customer positions can be represented by its centered counting function

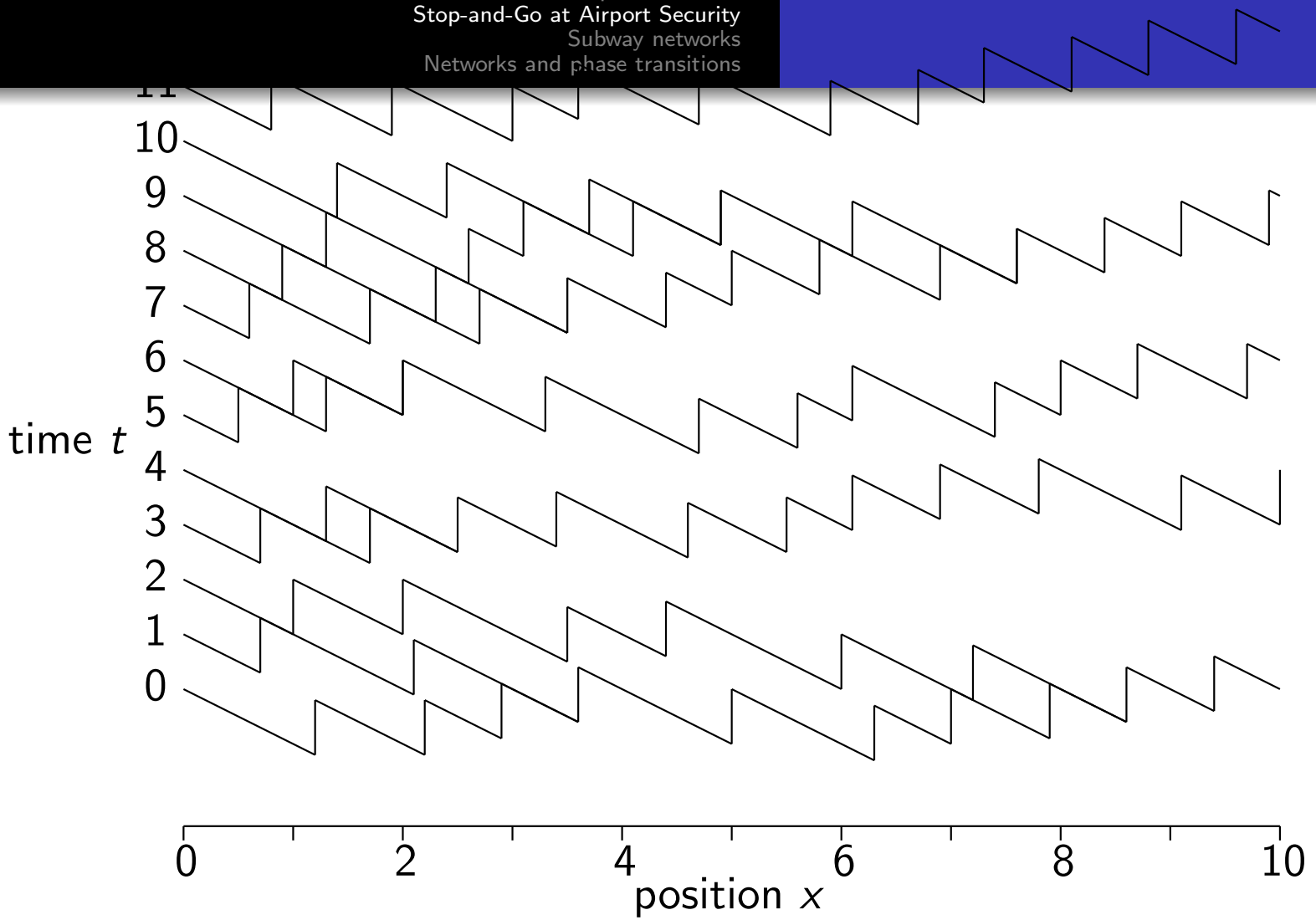
$$F(x) := \max\{k : \ x_k \le x\} \ - x, \quad 0 \le x < \infty. \tag{2}$$



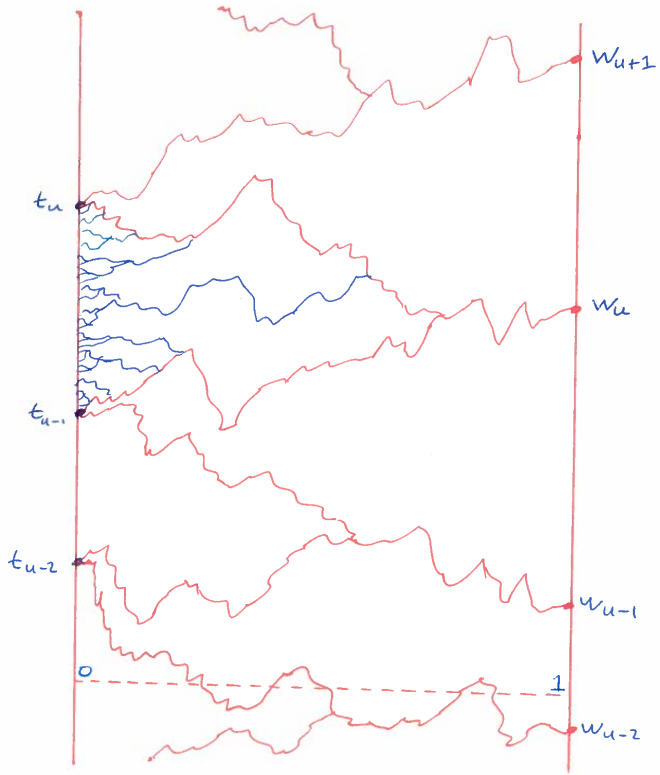
At each time t, let us consider the centered counting function $F_t(x)$ and plot the graph of the upward-translated function

$$x \to G(t,x) := t + F_t(x). \tag{3}$$

In other words, we draw the function starting at the point $(0, t)$ instead of the origin. Taking the same process realization as in the first Figure 1, superimposing all these graphs, gives the next Figure.



Why does this explain everything?



Picture shows **coalescing Brownian motion** (CBM) which is well understood; density of particles at time t scales as $t^{-1/2}$. But note trick: **we switched space** $\leftrightarrow$ **time**.

Assuming convergence of "coded" process to CBM, we can easily decode (details omitted) to get claimed "waves" result for the queue model.

How to actually prove the CBM limit?

AAAARGH !

30 pages with some details missing. Markov intuition fallible because of space $\leftrightarrow$ time switch. Must be some simpler proof ideas . . . . . .

Step -1.
Study W = length of wave at typical time. Suppose we can prove the desired order of magnitude

$$\mathbb{P}(W > w) \asymp w^{-1/2}.$$

Then we can lean on classical "random walk $\to_d$ BM" weak convergence, together with "robustness" of CBM – initial configuration unimportant for long-term behavior.

Topic 3: The shape of things to come?

In my "real world" context I describe typical math probability models (like the SIRSN) as "fantasy" – unconnected to any real data. But now I'll tell you an even more extreme fantasy – which will lead to an elementary-to-state math problem.

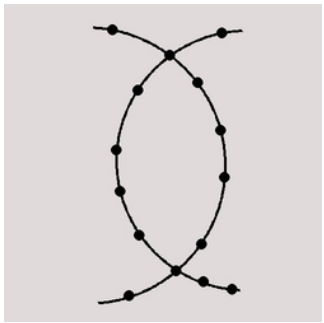
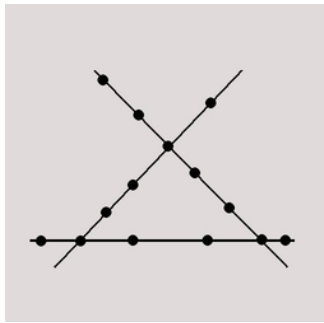
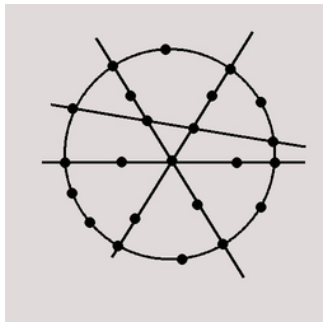
Imagine that somewhere there's an eccentric multi-billionaire with a taste for dramatic projects

and imagine there's a large spread-out metropolitan region without good public transport but with bad road traffic.

And then the billionaire has an idea . . . . . .

Where should one put a hypothetical such network?

Background. *Wikipedia – rapid transit* shows typical topologies (shapes) for subway-type networks.

An interesting problem – see Aldous-Barthelemy (2019) but not discussed today – is can we reproduce these as optimal under some slightly-realistic toy model?

Musk's hypothetical tunnel network suggests an extreme model: "infinite speed, no wait time, no discrete stations".

Problem (a) *Find the connected network of length L that minimizes the expected distance from a random start to the closest point on the network.*

This depends on the density ρ of starting point; as default take 2-dimensional standard Normal.

This model implies constant speed outside the network, infinite speed within the network, but one is forced to use the network. Slightly more realistic to allow a direct route outside the network:

Problem (b) Find the connected network of length L that minimizes the expected time $t(\xi_1, \xi_2)$ between independent(ρ) points,

$$t(\xi_1, \xi_2) = \min(||\xi_1 - \xi_2||, s(\xi_1) + s(\xi_2))$$

$$s(\xi) = \text{ distance from } \xi \text{ to the closest point on the network.}$$

(For large L the two problems are essentially the same).

- We seek the actual optimal network for each L – how does the shape evolve as L grows?
- We work numerically. Mostly we consider specific parameterized shapes and optimize over parameters
- Alternatively try simulated annealing to optimize over all networks.

Lemma: *An optimal network must be a tree (or single path).*

Because: If there is a circuit, removing a length ε segment costs order ε^2 but reattaching it elsewhere benefits order ε.

We do have a theorem concerning the $L \to \infty$ behavior. This result is not so interesting, so [recall name of Musk's tunneling company]

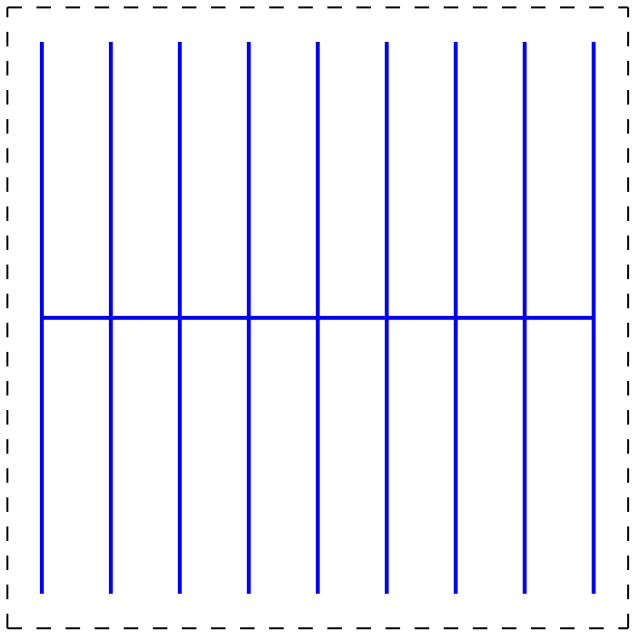
Take a starting density ρ. Write $d(L)$ for the expected distance-to-network in the optimal network of length L.

Theorem (The Boring Theorem)

$$d(L) \sim \frac{1}{4L} \left(\int_{\mathbb{R}^2} \rho^{1/2}(z)\, dz \right)^2 \quad \text{as } L \to \infty.$$

What the argument actually shows is that a sequence of networks is asymptotically optimal as $L \to \infty$ if and only if the rescaled local pattern around a typical position z consists of asymptotically parallel lines with spacing proportional to $\rho^{-1/2}(z)$, but the orientations can depend arbitrarily on z. Visualize a fingerprint.

Near-optimal network for uniform density on square.



A simple topology is the star network, with $n \geq 2$ branches of lengths L/n from the center, with optimal choice of $n = n_L$. Comparing with the other shapes we have examined leads us to the (rather unexciting)

Observation. For the Gaussian density, the star networks are optimal or near-optimal over the range $0 < L \leq 16$.

[We guess this is quite robust – true for other densities]

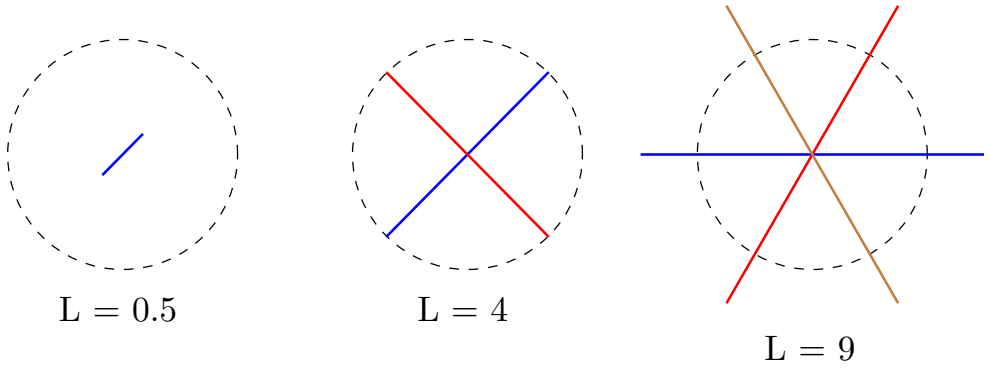


Figure 3: The optimal star networks for L = 0.5 and L = 4 and L = 9 (Gaussian density: the dashed circle indicates 1 s.d., so contains about 40% of the population).

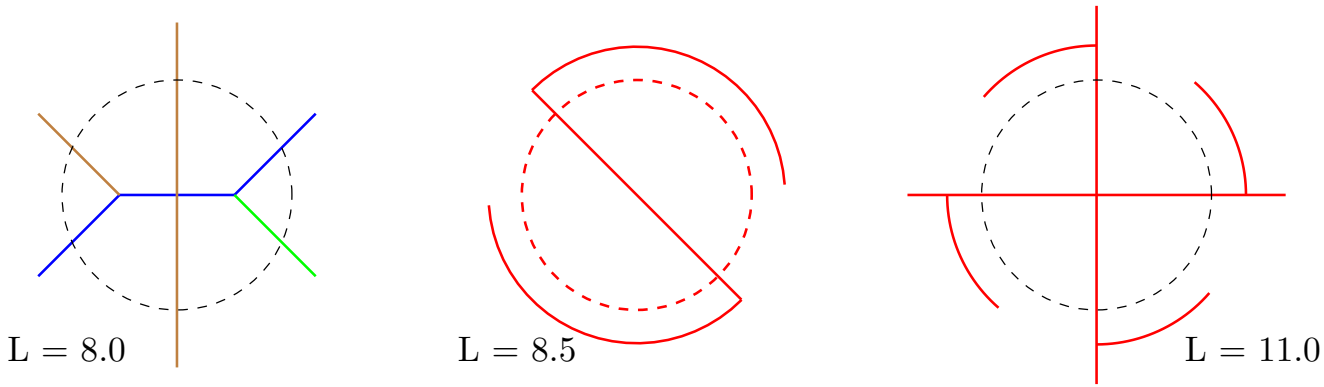


Figure 4: The "horse" network (left), the 2-arc network (center) and the 4-arc network (right).

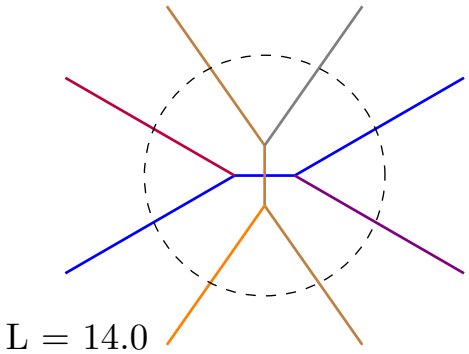


Figure 6: The spider network.

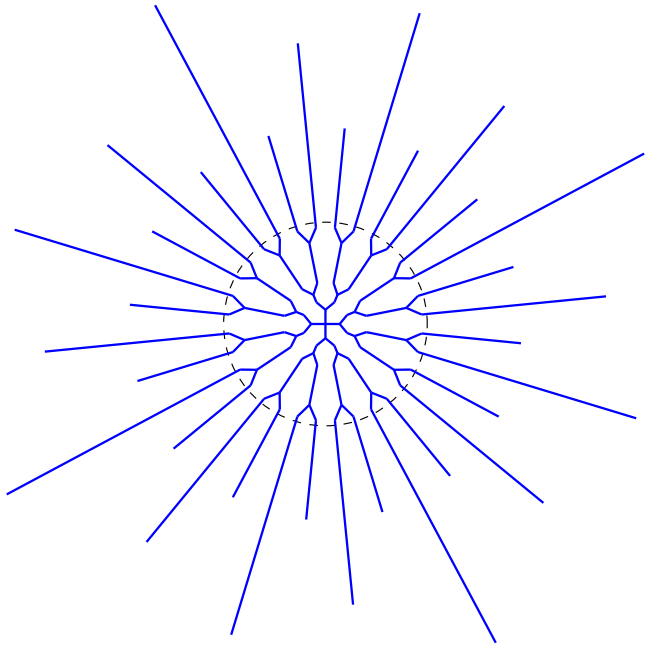


Figure 9: A finite approximation to an asymptotically optimal tree network.

As L grows an asymptotically optimal network becomes a branching tree.

Also one can construct **spirals** as asymptotically optimal. But contrary to our intuition, numerics say the tree is better (at second order).

Recall

Observation. For the Gaussian density, the star networks are optimal or near-optimal over the range $0 < L \leq 16$.

This was originally rather surprising.

By "reverse engineering" the Boring Theorem we see that star networks are asymptotically optimal for the non-Gaussian density of the rotationally invariant distribution on the radius-r_0 disc with R uniform on $[0, r_0]$.

Suggests robustness to density.

Conclusions from Setting 3: Optimal subway networks.

- Model is too unrealistic.
- Our intuition was poor.
- Don't hold your breath for the global sensation.

Topic 4: A Hard Open Research Problem.

To me a **network** is a finite (n vertices) connected edge-weighted undirected graph, vertices $v, x, y, \ldots$ and edge weights $w_e = w_{xy}$.

Note two opposite conventions for interpreting weights:

- In TSP-like setting, weight is distance or cost.
- In social networks, weight is strength of relationship (**this talk**).

Many stochastic processes can be defined over a general network. I will discuss bond percolation because it is essentially the SI epidemic model and I am interested in what one might be able to say about more realistic epidemic models.

Bond percolation.

An edge e of weight w_e becomes **open** *at an Exponential(w_e) random time.*

In this process we can consider

$$C(t) = \text{max size (number of vertices) in a connected component of open edges at time } t$$

This elates to "emergence of the giant component". Studied extensively on many non-random and specific models of random networks. Can we say anything about $n \to \infty$ asymptotics for (almost) arbitrary networks?

Suppose (after time-scaling) there exist constants $t_* > 0,\ t^* < \infty$ such that

$$\lim_n \mathbb{E} C_n(t_*)/n = 0; \quad \lim_n \mathbb{E} C_n(t^*)/n > 0. \tag{4}$$

In the language of random graphs, this condition says a *giant component* emerges (with non-vanishing probability) at some random time of order 1.

Proposition (1)

Given a sequence of networks satisfying (4), there exist constants τ_n such that, for every sequence $\varepsilon_n \downarrow 0$ sufficiently slowly, the random times

$$T_n := \inf\{t : C_n(t) \geq \varepsilon_n n\}$$

satisfy

$$T_n - \tau_n \to_p 0.$$

The Proposition asserts, informally, that the "incipient" time at which a giant component starts to emerge is deterministic to first order.

Reformulation as epidemics (well known but subtle).

An *SI* model refers to a model in which individuals are either *infected* or *susceptible*. In the network context, individuals are represented as vertices of an edge-weighted graph, and the model is

> *for each edge* (vy), *if at some time one individual* (v *or* y) *becomes infected while the other is susceptible, then the other will later become infected with some* transmission probability p_{vy}.

These transmission events are independent over edges. Regardless of details of the time for such transmissions to occur, this **SI model** is related to the **random graph model** defined by

> *edges* $e = (vy)$ *are present independently with probabilities* $p_e = p_{vy}$.

The relation is:

> *(*) The set of ultimately infected individuals in the SI model is, in the random graph model, the union of the connected components which contain initially infected individuals.*

In modeling an SI epidemic within a population with a given graph structure, we regard edge-weights $w_e = w_{vy}$ as indicating relative frequency of contact. Introduce a *virulence* parameter θ, and define transmission probabilities

$$p_e = 1 - \exp(-w_e\theta). \tag{5}$$

Note this allows completely arbitrary values of (p_e), by appropriate choice of (w_e). Now the point of the parametrization (5) is that the set of potential transmission edges is exactly the same as the time-θ configuration in the bond percolation model. So we can translate our Proposition into a statement about whether the SI epidemic model is pandemic (has $\Theta(n)$ vertices ultimately infected) in terms of the number κ_n of initially infected vertic

Even though this is mathematically trivial, it is **conceptually subtle**. A real-world flu epidemic proceeds in real-world time; instead we just consider the set of ultimately infected people and actual transmission edges; this structure, as a process parametrized by θ, is a nice stochastic process (bond percolation).

Proposition (2)

Take edge-weighted graphs with $n \to \infty$, consider the SI epidemics with transmission probabilities of form (5), and write $C'_{n,\kappa}(\theta)$ for the number of ultimately infected individuals in the epidemic started with κ uniformly random infected individuals. Suppose there exist some $0 < \theta_1 < \theta_2 < \infty$ such that, for all $\kappa_n \to \infty$ sufficiently slowly,

$$\lim_n n^{-1}\mathbb{E}C'_{n,\kappa_n}(\theta_1) = 0; \quad \liminf_n n^{-1}\mathbb{E}C'_{n,\kappa_n}(\theta_2) > 0. \tag{6}$$

Then there exist deterministic $\tau_n \in [\theta_1, \theta_2]$ such that, for all $\kappa_n \to \infty$ sufficiently slowly,

$$n^{-1}C'_{n,\kappa_n}(\tau_n - \delta) \to_p 0, \quad n^{-1}C'_{n,\kappa_n}(\tau_n + \delta) \gg_p 0$$

for all fixed $\delta > 0$.

Proposition 2 provides a subcritical/supercritical dichotomy for the SI epidemics under consideration. The conceptual point is that, for virulence parameter θ not close to the critical value τ_n, either almost all or almost none of the realizations of the epidemic affect a non-negligible proportion of the population. It really is a **phase transition**, and exists for essentially arbitrary large networks.

But the proof is very special. The **open problem** is to prove similar results for more realistic epidemic models.

References

- A Prediction Tournament Paradox. *The American Statistician*, 2019.
- Waves in a Spatial Queue. *Stochastic Systems*, 2017.
- (with Marc Barthelemy) The optimal geometry of transportation networks. *Physical Review E*, 2019.
- Limits for processes over general networks. *Talk slides, on my web site*.