Some Conceptual Issues Regarding Networks

David Aldous

June 2015

- Unfocussed talk – not addressing technical issues within the GraphEx world – seek to entertain with many digressions.
- As research I do mathematical probability, including graph/network models.
- I also have an ongoing "Probability in the Real World" project – I'll mention several aspects as we go. First, an undergraduate course in which I give 20 lectures on maximally different topics, each ideally anchored by some new data set of the type that students could obtain themselves.

- Everyday perception of chance
- Ranking and rating
- Risk to individuals: perception and reality
- Luck
- A glimpse at probability research: spatial networks on random points
- Prediction markets, fair games and martingales
- Science fiction meets science
- Coincidences, near misses and one-in-a-million chances.
- Psychology of probability: predictable irrationality
- Mixing: physical randomness, the local uniformity principle and card shuffling
- Game theory
- The Kelly criterion for favorable games: stock market investing for individuals
- Toy models in population genetics: some mathematical aspects of evolution
- Size-biasing, regression effect and dust-to-dust phenomena
- Toy models of human interaction: use and abuse
- Short/Medium term predictions in politics and economics
- Tipping points and phase transitions
- Coding and entropy

Here's how I start the "social networks" class; also it's the only math in my Public Lecture.

Why do your friends have more friends than you do, on average?

(This is the most interesting elementary math fact that most mathematicians don't know).

Simplest math formulation of a *social network* is as a graph: vertices are people, edges are a "friend" relationship. Write

$n =$ number of people
$d(i) =$ number of friends of person i, and $d(i) \geq 1$.
U (you) a uniform random person.
F (a typical friend) a uniform random friend of U.

Fact: $\mathbb{E}d(F) \geq \mathbb{E}d(U)$.

A fairly general point, in the context where you're trying to do something complicated, algorithmically, e.g. "community detection".

Conceptual Point

If you don't have an explicit "cost of error" function then you don't have a real-world problem.

For estimating a numerical quantity we are all familiar with MSE as mathematically convenient (thanks, Gauss). But when you think beyond mathematical convenience, MSE might be magically good (next slide) or may be a disaster (e.g. Netflix Prize).

Intend to present in a Public Lecture as paradox or magic trick – trying it out today on this audience, some of whom might know the trick.

I know how to give an exam (to students, say) with the following properties.

- There is a list of questions.
- I will grade the answers in an entirely objective way.
- The exam grades A, B, C will (for practical purposes) objectively represent students' relative abilities.
- No-one in the world knows the correct answer to any question.

(the questions involve the real world – nothing hypothetical or self-referential or Mathematical Fantasyland).

This sounds impossible – so what's the trick?

I participated in the IARPA-sponsored **Good Judgment Project** – volunteers in teams assess **probabilities** of future geopolitical events.

6 out of 65 questions asked several months ago: will the event happen **before 10 June 2015**?

- Will Russia officially lift its embargo on food imports from the US, the EU, Canada, Australia or Norway?
- Will a unity government be formed in Libya?
- Will the Ebola outbreak be contained in Liberia?
- Will there be a lethal confrontation between China's national military forces and the national military forces of another country in the South China Sea region?
- Will negotiations on the Transatlantic Trade and Investment Partnership be completed?
- Will the United Nations Security Council vote on the resolution referring the situation in North Korea to the International Criminal Court?

How can we judge someone's ability to assess probabilities of (for instance) future geopolitical events, where the true probabilities are unknown?

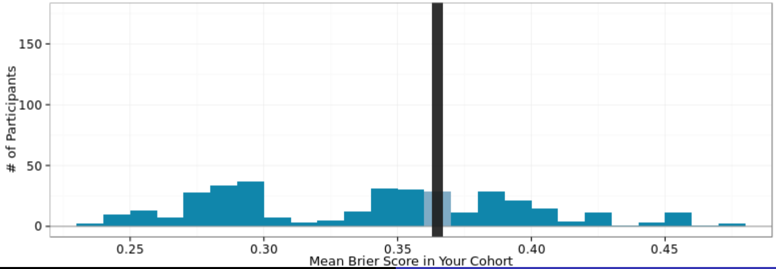
Answer: score by squared error. If you guess p then you score

- $(1-p)^2$ if event happens
- p^2 if not

and (like golf) you are trying to get a low score. Now

$$\mathbb{E}(score) = p_{true}(1-p_{true}) + (p_{est} - p_{true})^2.$$

Bottom line: One can assess people's abilities in relative terms, but not in absolute terms.



What I said above is mathematically elementary but conceptually subtle; it implicitly presupposes that (separately from mathematical convenience) $(p_{est} - p_{true})^2$ is a good way of measuring the cost of the error. This is not obvious, though I claim it is basically correct.

Claim. Outside of highly asymmetric settings (small chance of large loss/gain), in most contexts the cost of the error in estimating an unknown probability p_{true} is order $(p_{est} - p_{true})^2$ rather than $|p_{est} - p_{true}|$.

So freshman statistics – estimates from a size-n sample have error $O(1/\sqrt{n})$ – is true but misleading.

Obviously context-dependent but here are two contexts where we can justify this (separately from "mathematically convenient").

1. Simplest decision theory setup: choice of action 1 or action 2. If knew event A will happen then action 1 is preferable; if knew opposite then action 2 preferable. Unknown probability $p = \mathbb{P}(A)$.

- (Over ensemble of such decisions) frequency of wrong choices is $O(|p_{est} - p_{true}|)$.
- when you do make wrong choice, the utility cost is also $O(|p_{est} - p_{true}|)$.

2. Kelly criterion for favorable gambling. Suppose I offer to accept bets (either way) at fair odds based on my p_{est}; suppose you knew p_{true}. Then you can make money, but at what rate?

- (Over ensemble of such bets) amount you bet each time is $O(|p_{est} - p_{true}|)$.
- your mean gain per unit staked is also $O(|p_{est} - p_{true}|)$.

A network is a graph with context-dependent extra structure

Edges indicate 0/1 relationship. But I claim that in "most" real-world networks there is instead a quantifiable "strength of relationship", so one should start out as a default working with edge-weighted graphs.

Conceptual Point

The rapid development of a wide range of models (Erdos-Renyi, Strogatz-Watts, Albert-Barabasi, Chung-Lu, ...) of unweighted graphs, with focus on degree distribution, clustering etc, has distracted attention from this edge-weight feature.

When you are working with unweighted graphs, either what you're doing makes sense for weighted graphs or it doesn't. So

Conceptual Point

If it does make sense, use the edge-weights; if it doesn't, you're probably asking the wrong question.

In particular biological networks (cell biochemistry, gene regulation etc) are typically derived experimentally from measuring the level of some interaction, and then thresholding to give the graph.

Digression: I would like to be able to justify

claim: in "most" real-world networks there is a quantifiable "strength of relationship" that one can use to specify edge-weights.

To justify this claim one would want a well thought out list of 100 representative contexts where real-world networks arise; apparently no such list exists ......`so someone please create one!`

This exercise is in principle possible – I am doing this for Probability.

Topics in Nate Silver's *The Signal and the Noise*

- (3) accuracy of opinion polls vs expert assessments.
- (6) sports betting.
- (7) baseball players' performance.
- (8) professional poker.
- (16) flu pandemics.
- (26) weather.
- (32) mortgage default likelihoods.
- (34) predicting business cycle/economic indicators.
- (36) terrorism.
- (45) climate change.
- (65) predicting earthquakes.
- (81) stock market, efficient market hypothesis, bubbles.
- (84) herding, overconfidence.

I claim that in "most" real-world networks there is instead a "strength of relationship", so one should start out as a default working with edge-weighted graphs. The issue, of course, is that the interpretation of "strength of relationship" is highly context-dependent. In some contexts (scientific collaboration; corporate directorships) there is a natural quantitative measure, but not so in "friendship"-like contexts.

My particular research interest is to identify "strength of relationship" with "**frequency of meeting**", where "meeting" carries the implication of "opportunity to exchange information" rather than physical motion.

This puts us into some small sub-world within Network Science – my `desideratum` list of 100 would tell us how small.

Given finite connected graph with edge-weights ν_{ij}.

- Each pair i, j of agents with $\nu_{ij} > 0$ meets at random times, more precisely at the times of a rate-ν_{ij} Poisson process.
- An agent is in some "state"; when two agents meet they update their states according to some deterministic or random rule.

It turns out there is much theory fitting this setup, under names like "interacting particle systems" or "social dynamics" or "asynchronous random cellular automata". Many thousands of papers (originally in statistical physics) take particular update rules (e.g. "Voter Model") and particular graphs (e.g. $\mathbb{Z}^3$) and analyze theoretically how the model behaves.

Perhaps oddly, in most of this literature states are categorical not quantitative. In the *Voter Model* there are two possible states, say Red and Blue; when two agents with different states meet, one (random) agent changes state.

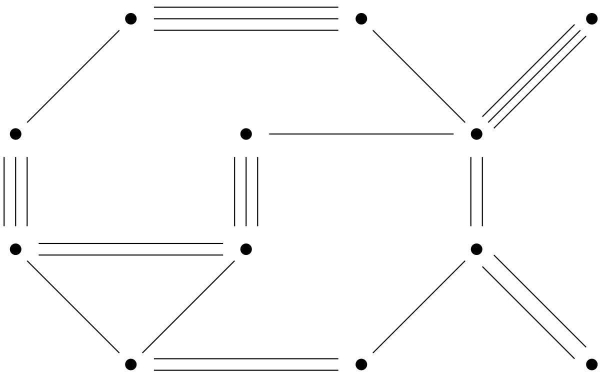
For fun I have studied two models where states are real numbers, envisaged as money. In the *Averaging Process*, agents start with different amounts; when two agents meet, they split their combined money equally. In the *Compulsive Gambler process*, agents start with the same amounts; when two agents meet, they make a fair-odds bet which results in one agent gaining all the combined money.

Here is one of my “theory” interests (not closely related to real-world data).

Suppose we don’t know the edge-weights (meeting rates) but can observe meetings for some time t. Suppose we’re interested in some process over the graph – we know the update rule – but we can’t observe states. And suppose **computation is free but observations are expensive**.

How well can we estimate some statistic of the process?

Data: numbers of observed meetings – can represent as multigraph.



Conceptually interesting to me because frequentist and Bayesian methodologies both seem reasonable but give different estimates.

A basic example is this class is the spectral gap λ of the graph Laplacian – that is, of the matrix (ν_{ij}) with $\nu_{ii} = -\sum_j \nu_{ij}$.

This arises because $1/\lambda$ is the *relaxation time* of the continuous-time Markov chain ("Hot Potato") with transition rates $= \nu_{ij}$, the edge-weights. The stationary distribution of this chain is always uniform (a convenient feature of the continuous-time chain).

Naive frequentist procedure:
estimate each rate ν_{ij} as $\widehat{\nu}_{ij} = N_{ij}(t)/t =$ observed rate of meetings;
estimate λ as $\widehat{\lambda}(t) =$ spectral gap of $(\widehat{\nu}_{ij})$.

Open Problem. Give a stopping rule T such that, observing up to time T, we have $\widehat{\lambda}(T)/\lambda$ is close to 1 with high probability (regardless of true edge-weights).

Coin-tossing analogy: to estimate $\mathbb{P}(Heads)$ so that $p_{est}/p_{true} \approx 1$ you toss until you have seen k Heads.

The spectral gap is one of many measures of "connectivity" for an edge-weighted graph. By coincidence I know a loosely related intriguing result relating to connectivity.

Theorem

Define T_k as the first time that our observed multigraph of meetings contains k edge-disjoint spanning trees. Then

$$\frac{s.d.(T_k)}{\mathbb{E} T_k} \leq \frac{1}{\sqrt{k}}.$$

Is there some analogous stopping rule for estimating spectral gap?

To do a Bayesian analysis one needs to put a prior on the space of all edge-weighted graphs on n vertices. This returns to a previous conceptual point: we want general-purpose models of edge-weighted graphs, in particular of sparse graphs.

I propose something to think about when designing such models – does it do better at estimating the spectral gap than the naive frequentist estimate?

The "naive Bayes" flat prior on each ν_{ij} is implicitly assuming (in the prior) the network extremely well connected, so will tend to over-estimate spectral gap. This is the opposite direction of error from naive frequentist, which estimates gap $= 0$ when the observed meeting multigraph is not connected.

Conceptual Point

Stochastic block models miss the point in community detection.

As a first step in comparing algorithms, OK, but way too far from realistic.

As well as previous issues (weighted edges, cost of errors) surely we often know in advance that certain types of communities will exist. The issue is whether there are other communities we don't know about. Analogous to multiple regression – to understand the influence of one factor, we need to discount the influence of other factors.

The notion of a "blank slate" algorithm – input a graph and output description of communities – makes little sense to me.