A framework for imperfectly observed networks

David Aldous

11 May 2016

A math model of a real-world network typically starts as a graph. This is weird, because almost all real networks are better represented as *edge-weighted* graphs. The reason this isn't the default (I guess) is that there are several conceptually different interpretations of edge-weight:

- flow capacity (road network, water network)
- distance or cost (TSP)
- strength of association (close friend or acquaintance or Facebook friend).

I'll consider the last class and think of *social networks* – collaboration networks, corporate directorships, Senators' voting record, etc (note many biological networks are also in this class). Even within this class of social networks there are different interpretations of *strength of association* , but (envisaging *friends*) I abstract this as *frequency of interaction*.

Introduce randomness by saying:

for each edge $e = (vy)$, individuals v and y interact at the times of a rate-w_e Poisson process.

So this is the meaning of the edge-weights $w_e \geq 0$.

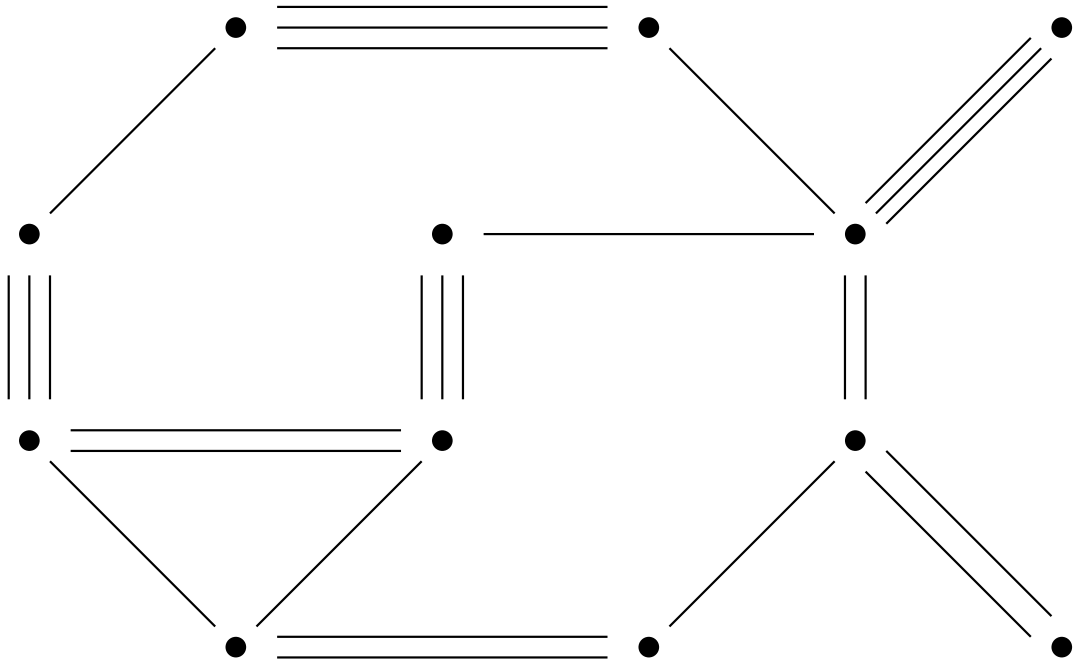
Aside. As discussed in my 2013 paper *Interacting Particle Systems (IPS) as Stochastic Social Dynamics* this setup underlies what probabilists call IPS: each individual is in some "state" and some update rule changes the states when individuals interact. This covers numerous models like the voter model or SIS epidemic – a line of research going back to statistical physics study of the Ising model on $\mathbb{Z}^d$.

This talk goes in a different direction: Suppose we are interested in some quantitative feature of a network which we could calculate if we knew exactly what the network is.
But suppose we don't know it . . . . . then what can we do?

I'll call this the **imperfectly-observed network** problem. I will talk about one particular formalization – not claimed to be useful for real-world data but (I do claim) interesting as math theory.

A **network** is a finite edge-weighted graph. We are concerned with some "statistic" Γ, a functional $G \to \Gamma(G)$ on finite edge-weighted graphs G. There is a network G^{true} with known vertices but unknown edges and edge-weights w_e. What we observe is the interaction process described above. That is, what we observe over time $[0, t]$ is the Poisson(tw_e) number of interactions $N_e(t)$ over edges e. We can represent our observations in two equivalent ways: either as the random multigraph with $N_e(t)$ copies of edge e, or as the random weighted graph $G^{\text{obs}}(t)$ in which edge e has weight $t^{-1}N_e(t)$.

How do we use these observations to estimate $\Gamma(G^{\text{true}})$, and how accurate is the estimate?



Some general comments.

- For any problem about networks where you assumed the network is known, you could ask this "imperfectly-observed" variation.
- There are many other ways to think about "imperfectly-observed networks" [one popular way will be shown later].
- We always have the naive frequentist estimator $\Gamma(G^{\rm obs}(t))$. It's natural to study, but there is no reason to think it is optimal.
- We always have the naive Bayes estimator (flat prior on each w_e) but .....
- "Computation is free" – not concerned with computational complexity – instead we regard observation time as the "cost".

Any estimator like $\Gamma(G^{\rm obs}(t))$ for fixed t will have error depending on the unknown $G^{\rm true}$. The "elegant" formulation of a mathematical problem is:

Program

Given a statistic Γ*, define a ("universal") stopping rule* T *and an estimator such that the relative error of the estimator, say* $\Gamma(G^{\rm obs}(T))/\Gamma(G^{\rm true}) - 1$*, is w.h.p. small* **uniformly** *over all networks* $G^{\rm true}$*.*

Program

Given a statistic Γ, *define a ("universal") stopping rule* T *and an estimator such that the relative error of the estimator, say* $\Gamma(G^{\rm obs}(T))/\Gamma(G^{\rm true}) - 1$, *is small* **uniformly** *over all networks* $G^{\rm true}$.

The bottom line of this talk. We have no idea how to do this for most interesting/natural statistics, but we can do this for a few statistics which are less interesting/natural.

This is ongoing joint work with grad student Lisha Li.

Given $(G, \mathbf{w})$, write n for the number of vertices and $w_v = \sum_y w_{vy}$ for the total interaction rate of vertex v. We are thinking of results for large networks, formalized as $n \to \infty$ limits. **For discussion purposes here** (not as assumptions in theorems) assume $w_v \equiv 1$, so in time t we have seen on average t interactions involving each vertex, that is our observed multigraph has on average t edges at each vertex.

Qualitatively there are 3 time regimes.

- For $t = o(1)$ can only estimate statistics like (weighted) degree distributions (cf. birthday problem).
- To make the observed graph connected we typically need $t = \Theta(\log n)$ (cf. coupon collector problem) at which time we see $\Theta(\log n)$ edges per vertex and (intuitively) "we can estimate anything well".
- The interesting/challenging regime is where t is a (large-ish) constant; what can we infer when we have seen 33 interactions per individual?

The weird logic of freshman (frequentist) statistics

Suppose we have a theorem of the format

> **Theorem:** *if* G^{true} *has property* Q^* *then with* $\geq 95\%$ *probability* G^{obs} *has property* Q.

We can restate this as an inference procedure of the format

> **Inference:** *if* G^{obs} *does not have property* Q *then we are* $\geq 95\%$ *confident that* G^{true} *does not have property* Q^*.

But we want to state the inference in "positive" terms, so we negate the property and restate as follows.

If we wish to justify an inference procedure of the format

> **Inference:** *if* $G^{\rm obs}$ *has property* P *then we are* $\geq 95\%$ *confident that* $G^{\rm true}$ *has property* P^*

then we need to prove a theorem of the format

> **Theorem:** *if* $G^{\rm true}$ *does not have property* P^* *then with* $\geq 95\%$ *probability* $G^{\rm obs}$ *does not have property* P.

Usually with random graph models we are interested in establishing some "desirable" property; paradoxically in our framework we need to show $G^{\rm obs}$ has "worse" properties than $G^{\rm true}$. But our intuition is that the randomness in $G^{\rm obs}$ will typically make it "worse" than $G^{\rm true}$, so this might be true.

For a first concrete statistic, recall that connectedness of a weighted graph (G or $\mathbf{w}$) is often quantified by the *spectral gap of the graph Laplacian*, that is of the symmetric matrix $\mathbf{w}$ extended to the diagonal via

$$w_{vv} = -w_v = -\sum_{y \neq v} w_{vy}.$$

It is immediate from the extremal characterization of spectral gap that

$$\mathbb{E}\,\mathrm{gap}(G^{\mathrm{obs}}(t)) \le \mathrm{gap}(G^{\mathrm{true}}).$$

The inequality goes in the right direction, but may be trivial: the gap is zero while $G^{\mathrm{obs}}(t)$ is not connected. Better to use some stopping time T at which G^{obs} is connected. Note

$$\mathbb{E}\,\mathrm{gap}(G^{\mathrm{obs}}(T)) \le \mathrm{gap}(G^{\mathrm{true}}) \quad \textbf{FALSE}$$

Need to use un-normalized edge-counts $\mathbf{N}(t) = (N_e(t))$ to get

$$\mathbb{E}\,\mathrm{gap}(\mathbf{N}(T)) \le \mathrm{gap}(G^{\mathrm{true}})\ \mathbb{E}\,T.$$

Still only half-satisfactory – running the observation process once gives one realization of the pair $(T, \mathrm{gap}(\mathbf{N}(T)))$ but would need to repeat the process, **unless** we knew these RVs are concentrated around their means.

Alas, connectivity (above) involves the wrong time regime $t = \Theta(\log n)$. Here is a fundamental, albeit vague, open problem in the "interesting" time regime $t = \Theta(1)$.

> *if we observe $G^{\rm obs}(t)$ has a "highly connected" (in some sense) giant vertex set of size αn, then we can infer that $G^{\rm true}$ has a similarly "highly connected" giant vertex set of size $\beta(\alpha)n$?*

There are many ways to quantify connectedness by a statistic Γ in this context, for instance via spectral gap of the (restricted) graph Laplacian. We conjecture that our program (repeated below) can be done in this setting. The *intuition* is that randomness makes $G^{\rm obs}$ *less* well connected than $G^{\rm true}$ – but we have no idea how to prove any reasonable version.

Program

Given a statistic Γ, define a ("universal") stopping rule T and an estimator such that the relative error of the estimator, say $\Gamma(G^{\rm obs}(T))/\Gamma(G^{\rm true}) - 1$, is small **uniformly** *over all networks $G^{\rm true}$.*

On the positive side, here is a "sideways" approach to our program. Consider

$$T_k^{tria} = \inf\{t : \text{ observed multigraph contains } k \text{ edge-disjoint triangles}\}.$$

$$T_k^{span} = \inf\{t : \text{ observed multigraph contains } k \text{ edge-disjoint spanning trees}\}.$$

Proposition

$$\frac{\text{s.d.}(T_k^{tria})}{\mathbb{E}\, T_k^{tria}} \le \left(\frac{e}{e-1}\right)^{1/2} k^{-1/6}, \ k \ge 1.$$

$$\frac{\text{s.d.}(T_k^{span})}{\mathbb{E}\, T_k^{span}} \le k^{-1/2}, \ k \ge 1.$$

So here the bounds are independent of $\mathbf{w}$, meaning that we can estimate the statistics $\mathbb{E}\, T_k$ without assumptions on $\mathbf{w}$.

So the "sideways" approach is to seek some observable quantity which is concentrated around its mean, independent of $\mathbf{w}$, which therefore provides an estimator of the statistic defined by the expectation.

From arXiv preprint *Weak Concentration for First Passage Percolation Times on Graphs and General Increasing Set-valued Processes* and the title give a hint of the proof method.

Our observation process, considered as a growing multigraph, is an increasing set-valued process, for which there is a simple general bound on $\frac{\text{s.d.}(T)}{\mathbb{E}T}$ for the first time T that some "increasing" property holds. In our context, we have

$$T_k = \inf\{t : \text{ observed multigraph contains } k \text{ edge-disjoint } \textbf{objects}\}$$

and the argument for the bound uses only one object-specific calculation, which I will outline as a game, which is trivial in the two cases (triangles and spanning trees) above.

The game. I choose a multigraph with the given "contains k edge-disjoint **objects**" property, and I then delete an edge, and then show you. Can you always find many different ways to restore the property by creating a few new edges?

Spanning trees; deleting edge creates a split $(A, \mathcal{V} \setminus A)$ of vertex-set $\mathcal{V}$; sufficient for you to create any edge between A and $\mathcal{V} \setminus A$.

Triangles: sufficient for you to create one new triangle.

The bound in the general inequality involves (worst-case) mean "restore" time in the observation process.

Open problem; Can we do this for the "k-edge connected" property? (Menger's theorem doesn't seem to help).

Here is a first example of a "natural" statistic. Identify a graph with its matrix $\mathbf{w}$ of edge-weights.

Maximum matching. Take n even. A *matching* is a set π of $n/2$ edges such that each vertex is in exactly one edge.
The *weight* of the matching is $\text{weight}(\pi, \mathbf{w}) := \sum_{e \in \pi} w_e$.
The *maximum-weight* is $\Gamma_1(\mathbf{w}) := \max_\pi \text{weight}(\pi, \mathbf{w})$.
Can we estimate $\Gamma_1(\mathbf{w})$ from the observed $G^{\text{obs}}(t)$ at (large) times $t = O(1)$?

The naive frequentist estimator $\Gamma_1(G^{\text{obs}}(t))$ does not work – consider the "dense" case of the complete graph with edge-weights $w_e = 1/(n-1)$.

We will finesse this issue by reformulating the problem. Because real-world networks are typically sparse, we can say that, although we require our estimates to be **valid** for all G^{true}, we only require them to be **informative** for sparse G^{true}.

Informally, we regard a weighted graph as *sparse* if the vertex-weight sums $w_v = \sum_y w_{vy}$ are dominated by the largest $O(1)$ terms.

For discussion, assume $w_v \equiv 1$. For a sparse graph we will have $\Gamma_1(\mathbf{w}) = \Theta(n)$, so we reformulate the problem as

can we estimate $n^{-1}\Gamma_1(\mathbf{w})$ *up to small additive error?*

Such an estimator will be informative in the sparse case, but not for *dense* graphs like the complete graph above, for which $\Gamma_1 = \Theta(1)$.

A moment's thought says that to know anything about the weight of some specific edge we must observe at least two interactions (cf. unseen species problem).

This suggests making an estimator using only edges for which we have observed at least two "interactions". That is, we define

$$\text{weight}_2(\pi, G^{\text{obs}}(t)) := t^{-1} \sum_{e \in \pi} N_e(t) 1_{\{N_e(t) \geq 2\}}$$

$$\Gamma_2(G^{\text{obs}}(t)) := \max_{\pi} \text{weight}_2(\pi, G^{\text{obs}}(t))$$

and our goal is to obtain a bound of the form

$$\mathbb{E} n^{-1} \left| \Gamma_2(G^{\text{obs}}(t)) - \Gamma_1(\mathbf{w}) \right| \leq \psi(t) \ \forall \mathbf{w}. \tag{1}$$

The best we can hope for is a $\psi(t) = O(t^{-1/2})$ bound: consider the graph with only one edge. And a conceptually straightforward argument (large deviations and counting) shows (1) is true for some

$$\psi(t) = O(t^{-1/2} \log t).$$

[Also a factor $\max_v w_v$, but we can estimate this more quickly].

Observed and true community structure.

For a subset A of vertices write A^* for the set of edges with both end-vertices in A. Write

$$\overline{\mathbf{w}}_m^{\text{true}} = m^{-2} \max \left\{ \sum_{e \in A^*} w_e \ : \ |A| = m \right\}$$

– essentially the maximum edge-density in a size-m community. Ignoring computational complexity, suppose we can compute the analogous observable quantity

$$\overline{W}_m^{\text{obs}}(t) = m^{-2} \max \left\{ \sum_{e \in A^*} N_e(t)/t \ : \ |A| = m \right\}.$$

To make inferences from the observed $G^{\text{obs}}(t)$ to G^{true} we need $m \sim \gamma \log n$. Then (as in previous example, just using large deviations and counting) we can be confident that $\overline{\mathbf{w}}_m^{\text{true}}$ is in a certain interval, roughly

$$\left[\overline{W}_m^{\text{obs}}(t) - \sqrt{\frac{2\overline{W}_m^{\text{obs}}(t)}{\gamma t}}, \overline{W}_m^{\text{obs}}(t) \right].$$

Here is one case where it seems **impossible** to carry out this program. It is a basic example of a process built over a weighted graph.

First passage percolation (FPP).

(somewhat different setting from Remco's]

Given an edge-weighted graph $(G, \mathbf{w})$ with distinguished vertices (v^*, v^{**}), create independent random variables ξ_e with Exponential(w_e) distributions, and view ξ_e as the "traversal time" of edge e. Let $X(\mathbf{w})$ be the (random) FPP time from v^* to v^{**}, that is the minimum value of $\sum_{e \in \pi} \xi_e$ over all paths π from v^* to v^{**}. Take the expectation of this FPP time as our statistic

$$\Gamma(\mathbf{w}) = \mathbb{E}X(\mathbf{w}).$$

First 2 observations

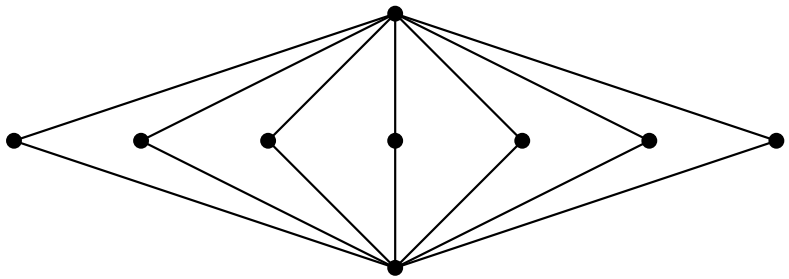
- For any G we can estimate $\Gamma(\mathbf{w})$ roughly (order of magnitude) in time $\Gamma(\mathbf{w})$ by using the observation process to simulate the FPP process itself.
- For a linear graph, edge-weights unknown but $\Theta(1)$, we have $\Gamma = \Theta(n)$ but we can estimate in observation time $\Theta(\log n)$.

We spent some effort trying to find

> *a universal estimator of* $\Gamma(\mathbf{w})$ *whose observation time* T *is always* $O(\Gamma(\mathbf{w}))$ *and for some "nice" graphs is* $o(\Gamma(\mathbf{w}))$.

But the following construction convinces us this is impossible!

There exist graphs G^* where "observation time needed" and actual FPP time are the same order.



So given a "nice" graph with $T \ll \Gamma(\mathbf{w})$ we could superimpose such a graph G^* whose times were inbetween those times.. This will fool the algorithm.

3 more 1-slide topics, in different directions.

- A very different framework for "imperfectly-observed networks".
- What about Bayes?
- Bond percolation on general weighted graphs.

A very different framework for "imperfectly-observed networks".

[from a 2011 survey *Link prediction in complex networks* by Linyuan Lü and Tao Zhou, cited 683 times.]

Consider unweighted graphs, and only the possibility of unobserved edges – this is called *link prediction*. In this literature, the goal is to define an algorithm that takes the observed edges as input, and outputs an ordering $e_1, e_2, \ldots$ of all the other possible edges, intended as decreasing order of assessed "likelihood" of the edge being present. This is done by defining, for each possible edge (v_1, v_2), some statistic based on (typically) the local structure of the observed graph near v_1 and v_2, for instance

$$s(v_1, v_2) = \frac{|\mathcal{N}(v_1) \cap \mathcal{N}(v_2)|}{|\mathcal{N}(v_1)|\ |\mathcal{N}(v_2)|}$$

where $\mathcal{N}(v)$ is the set of neighbors of v. Then list edges in decreasing. order of $s(v_1, v_2)$.

In this framework there is no probability model involved; different algorithms are compared empirically by taking a real-world network, randomly deleting a proportion of edges to create a synthetic "observed graph", and comparing the algorithms' effectiveness in predicting the deleted edges.

Bayesian approach. Returning to our framework – unknown $G^{\rm true}$ and an observed $G^{\rm obs}(t)$ – it is conceptually simpler to take the Bayesian view. Put a prior on $G^{\rm true}$, compute the posterior distribution of $G^{\rm true}$ given $G^{\rm obs}(t)$, then any given statistic has a posterior distribution.

In particular, if we assume $G^{\rm true}$ is connected and wish to estimate the spectral gap of the graph Laplacian, in our previous setup we need $t = \Theta(\log n)$ to make $G^{\rm obs}(t)$ connected and get a non-trivial estimate, where in the Bayes setup we can put a prior on connected graphs.

But not so easy in practice – how do you choose a plausible prior?

To play with the mathematics, consider the "naive Bayes" procedure – take as prior the uniform law on $[0, \infty)$ for each w_{ij} – for which the posterior distribution on $\mathbf{w}$ given observed interactions (n_{ij}) is that the w_{ij} are independent with densities

$$\nu \to p(n_{ij}; \nu t) \tag{2}$$

where $p(k; \lambda)$ denotes the Poisson probability function.

Informally, this "flat" prior lives on highly connected graphs, and for small t the posterior distribution on $\mathbf{w}$ will concentrate on too-highly-connected graphs, with spectral gap around ne^{-t}. So we will not get a good estimate of true spectral gap before time $\Theta(\log n)$.

Bond percolation and giant components.

Take our background setting of an arbitrary edge-weighted n-vertex graph $(G, \mathbf{w})$. To the edges $e \in \mathbf{E}$ attach independent Exponential(rate w_e) random variables ξ_e. In the language of percolation theory, say that edge e becomes *open* at time ξ_e. The set of open edges at time t determines a random partition of vertices into connected components; write $C(t)$ for the largest number of vertices in any such connected component.

Next result from arXiv preprint *The Incipient Giant Component in Bond Percolation on General Finite Weighted Graphs*.

Now consider a sequence of such weighted graphs with $n \to \infty$, where both the graph topologies and the edge-weights are arbitrary subject only to the conditions that for some $0 < t_1 < t_2 < \infty$

$$\lim_n \mathbb{E} C_n(t_1)/n = 0; \quad \lim_n \mathbb{E} C_n(t_2)/n > 0. \tag{3}$$

In the language of random graphs, this condition says a *giant component* emerges (with non-vanishing probability) sometime between t_1 and t_2.

Proposition

Given a sequence of graphs satisfying (3), there exists a deterministic sequence $\tau_n \in [t_1, t_2]$ 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

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

Proposition 2 asserts, informally, that the "incipient" time at which the giant component starts to emerge is deterministic to first order.