A framework for imperfectly observed networks

David Aldous

12 July 2016

The talks at this workshop cover a range of topics which is very broad, but loosely form two categories.

- Properties of specific probability models of random graphs
- Algorithms/statistical estimation for problems over arbitrary graphs.

This talk is midway between. Envisage an arbitrary true network we can't observe, and devise a probability model for observed "noisy" network. How do we estimate some statistic – some quantitative feature – of the true network?

Aside. There are many other ways to model "imperfectly observed networks" – e.g. talks by Peter Orbanz and by Elizaveta Levina. My formulation is not claimed to be very useful for real-world data but (I do claim) interesting as math theory.

Rant # 17

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*.

A **network** is a finite edge-weighted graph intended to model something in the real world. In contexts where edge-weights w_e indicate some notion of *strength of association* it is reasonable to assume that stronger associations are easier to observe.

One way to quantify *strength of association* is to interpret it as *frequency of interaction* and to suppose what we observe is the interactions. This suggests a probability model:

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

and we observe these interactions.

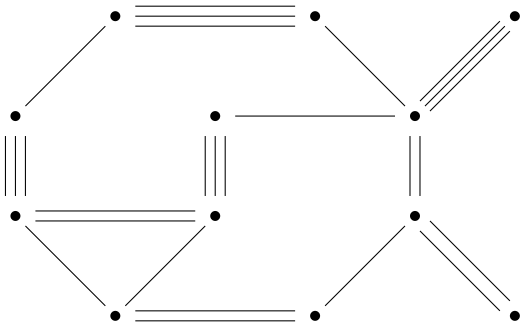
That is, what we observe over time $[0, t]$ is the number $N_e(t)$ of interactions over edges e.

So this provides our *framework for imperfectly observed networks*. To repeat in different words ....

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 $G^{\mathrm{obs}}(t)$ in which edge e has weight $t^{-1}N_e(t)$.

How do we use these observations to estimate $\Gamma(G^{\mathrm{true}})$, 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.
- We always have the naive frequentist estimator $\Gamma(G^{\mathrm{obs}}(t))$. 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 $\Gamma(G^{\mathrm{obs}}(t))$ for fixed t will have error depending on the unknown G^{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 $\Gamma(G^{\mathrm{obs}}(T))/\Gamma(G^{\mathrm{true}}) - 1$, is w.h.p. small **uniformly** *over all networks G^{true}.*

Program

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

This is ongoing joint work with grad student Lisha Li.

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.

The rest of this talk:

- A typical "easy" example.
- A key open problem.
- A backwards approach.

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

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

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

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

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

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

A similar but more complicated estimator works for **max-weight matching**.

Let's standardize time until so that there are $O(1)$ observed interactions per vertex per unit time. The estimators above require only $O(1)$ time – this is "the interesting case".

Informally, what can we say about G^{true} when we have observed an average 24 interactions per vertex?

A key open problem.

Typically G^{true} will be connected; but (coupon collector) typically we need order $\log n$ observed interactions per vertex until G^{obs} is connected – "not the interesting case" because then we can estimate almost all w_e accurately.

Here is a fundamental, albeit vague, open problem in the "interesting" time regime $t = \Theta(1)$.

> *if we observe $G^{\mathrm{obs}}(t)$ has a "highly connected" (in some sense) giant vertex set of size $(1-\delta)n$, then we can infer that G^{true} has a similarly "highly connected" giant vertex set of size $(1-\beta(\delta))n$?*

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

Digression: proving inference assertion involves

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 instance in the "well-connected very large component" context above).

On the positive side, here is a "backwards" approach to our program, illustrated by example. 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

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

$$\frac{\mathrm{s.d.}(T_k^{span})}{\mathbb{E}\, T_k^{span}} \leq k^{-1/2}, \ k \geq 1.$$

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

So the "backwards" approach is to seek some observable quantity which is concentrated around its mean, independent of **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 $\frac{\mathrm{s.d.}(T)}{\mathbb{E} T}$ 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).