A general SI epidemic and a framework for imperfectly observed networks

David Aldous

6 July 2016

Talk is rather off-topic.

- I have some “big picture” projects about networks.
- One rather fringe result has an SI epidemic interpretation.
- This suggests conjecture for (very general) SIR epidemics.

Key thought; Can we say anything at all about SIR epidemics on networks without assuming some specific network model (configuration model, etc)?

We all believe some version of the following:

Given some probability model with a bunch of parameters for an epidemic in a (large) size-n population:

- for some subset of parameter space, the epidemic process is subcritical: number infected is $o(n)$ with high probability
- for another subset of parameter space, the epidemic process is supercritical: number infected is not $o(n)$, with probability not too close to zero
- the remaining " critical" subset of parameters is small.

Why do we believe this?

- R_0 is some function of the parameters.

In what generality can we prove it?

Big picture background

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 $w_e \geq 0$.

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 SIR/SIS epidemic – a line of research going back to statistical physics study of the Ising model on $\mathbb{Z}^d$.

Within this huge field, to get explicit results one needs an explicit model for the network. But mathematically tractable models are usually unrealistic. Is there any hope to say something about a given IPS rule over an arbitrary finite network? I will describe a theorem, not itself so relevant to epidemics, but suggestive of a very general result for epidemics.

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 $0 < t_1 < t_2 < \infty$

$$\lim_n \mathbb{E} C_n(t_1)/n = 0; \quad \lim_n \mathbb{E} C_n(t_2)/n > 0. \tag{1}$$

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 (1), 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 1 asserts, informally, that the "incipient" time at which the giant component starts to emerge is deterministic to first order.

Even though the most natural formulation is via the dynamic random graph – a generalization of the Erdos-Renyi process over time $0 < t < \infty$ - -we can reformulate the result in terms of SI epidemics. Regard t as a fixed parameter. Model

- Infection will spread across an undirected edge vy with probability $1 - \exp(-tw_{vy})$.

Then the ultimately infected set, for an SI epidemic started at v_0, is just the component of the random graph process at t containing v_0. We can reformulate the Proposition as a subcritical/supercritical dichotomy for this SI epidemic, as follows.

- Infection will spread across an undirected edge vy with probability $1 - \exp(-tw_{vy})$.

Take a number $\omega_n \uparrow \infty$ arbitrarily slowly of random initial infectous, and let $C'_n(t)$ be the total size of the SI epidemic. Take arbitrary networks, subject to

$$\lim_n \mathbb{E} C'_n(t_1)/n = 0; \quad \lim_n \mathbb{E} C'_n(t_2)/n > 0. \tag{2}$$

for some $0 < t_1 < t_2 < \infty$.

Proposition

Then there exist deterministic $\tau_n \in [t_1, t_2]$ such that, with probability $\to 1$,

$$C'_n(t) = o(n) \text{ for } t \le \tau_n - \varepsilon$$

$$C'_n(t) \neq o(n) \text{ for } t \ge \tau_n + \varepsilon$$

So behavior is same for most realizations, depending on parameter t. My proof leans on having Exponential distributions but (intuitively) must hold much more generally.

Wild vague conjecture

Define a set $\mathcal{H}$ of distribution functions, for some family of distributions "not wildly different from Exponential" (or zero function).

Introduce a virulence parameter θ.

Model an SIR epidemic: for population size n assume

- infectous duration for v has distribution function $\iota_v(\theta)$ in $\mathcal{H}$, and is stochastically increasing in θ.
- Infection will spread across an undirected edge vy with probability $p_{vw}(\theta)$, where function $\theta \to p_{vw}(\theta)$ is in $\mathcal{H}$.

Take a number $\omega_n \uparrow \infty$ arbitrarily slowly of random initial invectives, and let $C'_n(\theta)$ be the final size of the SIR epidemic. Take arbitrary networks and parameters, subject to distributional assumptions above and

$$\lim_n \mathbb{E} C'_n(\theta_1)/n = 0; \quad \lim_n \mathbb{E} C'_n(\theta_2)/n > 0. \tag{3}$$

for some $0 < \theta_1 < \theta_2 < \infty$.

Conjecture. Then there exist deterministic $\theta^*_n \in [\theta_1, \theta_2]$ such that, with probability $\to 1$,

$$C'_n(\theta) = o(n) \text{ for } \theta \leq \theta^*_n - \varepsilon$$

$$C'_n(\theta) \neq o(n) \text{ for } \theta \geq \theta^*_n + \varepsilon.$$

How was the original Proposition proved?

I'm not telling you here! (read the arXiv preprint).

Does not use any notion of R_0.

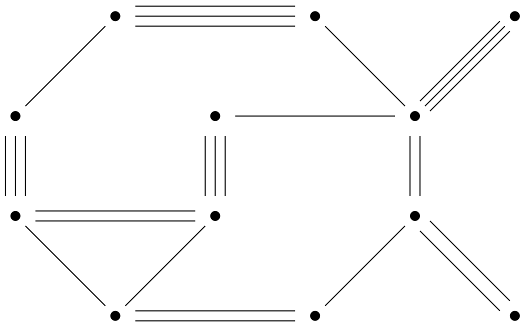
The background project

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 at the start of this talk. 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.
- 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 $\Gamma(G^{\mathrm{obs}}(t))$. 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}*.*

The bottom line of this project. 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 an average of 24 interactions per individual?

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

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

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

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 "sideways" 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.

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

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

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

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

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

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