12/4/22 Previously, I was trying to work up my own estimate for the probability of the extinction time being τ, $\hat{p}_i(\tau)$. But then, I realized I'd have to account for the messages that haven't gone extinct by the time recording ends.

Fortunately, this is very similar to the problems of survival analysis; hence, I've been considering if the Kaplan-Meier curve is appropriate for this problem.

1) The Kaplan-Meier (KM) curve estimates the survival function

$$S(\tau) = \mathbb{P}(T > t),$$

but this is clearly related to the $p_i(t)$ values I'm interested in by

$$p_i(t) = S_i(t) - S_i(t-1) \quad i \in [n]$$

where S_i is the survival function for messages injected at node i. <u>Therefore, the KM curve does address my problem.</u>

2) There are certain assumptions underlying the KM estimator, and it's important to address these assumptions. I've found a starting point for this issue

[1] Piovani, Nikolopoulos, Bonovas: "Pitfalls and perils of survival analysis..." (2021);

we'll return to this after I write my last consideration.

3) The KM curve is an estimator of the survival function, so if we can satisfy what assumptions come out of consideration 2, then our next concern would be with the convergence properties of this estimator:
- Is it biased?
- Is it consistent?
- What is the rate of convergence? This question is crucial, as the answer will tell us how many messages we need to inject at each node

Back to consideration 2 now, we're taking notes from Piovani et al

Independent Censoring a) Censoring should be "non-informative"; patients who are censored should have the same risk of event occurence, conditional on exposure and covariates (what does this mean?) as those who continue to be followed. This is called the "assumption of independent censoring." Violation can bias the KM estimate.

Proportionality b) We often want to compare survival between two or more groups of patients (eg different treatments or different baseline characteristics). The Cox proportional hazards model is the standard tool for this. A fundamental assumption is that the ratio of hazards of any two individuals should remain roughly constant over time. (the proportionality assumption)

- Base assumption for log-rank test & proportional Cox regression method.
- KM curves between groups would be roughly parallel functions
- This is something you have to test (Schoenfeld residuals are one method).
- If the proportionality assumption not met ⇒ estimated hazards ratios depend on follow-up time (i.e., observation period)

c) Competing risks: when the chance to observe the outcome of interest can be altered or prevented by the occurrence of a competing event. KM estimates are not appropriate in this case.

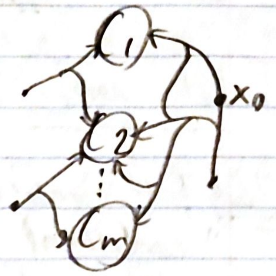
So in this problem, competing risks are not a problem at all, since extinction of a message can only occur due to the activity on the network.

Proportionality is something we can't assess until we get the data, but I have a feeling that this depends on the $\boxed{\text{Markov chain}}$ structure. To elaborate, recall that for each digraph A and alphabet B, the ~~dynamical process~~ Markov chain

I never proved that this is a Markov chain, it just seemed obvious to me. Should check

$$x_{t+1} = f(Ax_t + b_t), \quad x_0 \sim unif([1]_0^n)$$

gives rise to a digraph on X_n. Say this digraph has communicating classes $C_1, \ldots, C_m$ and also some transient states, and then say x_0 is a transient state.



If we start monitoring extinction times from $t=0$, then the extinction times may change as $t \to \infty$ since the state of the Markov chain will be absorbed by a communicating class.

The ergodic case is best because over time, we would explore one communicating class, so the probability of extinction would converge. But telling for which A and B this chain is ergodic is really hard, and in any case, we're not ultimately interested in restricting to the ergodic case.

Multiple trials from the same starting state would reach each communicating class available to that state but carrying this out via simulation ~~would be~~ is unfeasible.

X One hack I've been thinking of is to modify the alphabet by including I, but assigning it very small probability. This would make the Markov chain ergodic ... <u>FALSE</u>

We can save this idea. Fix $L \in [n]_0$, $0 < \varepsilon < 1$. The alphabet B would be $[1]_0^n$, but the probability distribution would be like $(1-\varepsilon)\,unif(B_L)$ over B_L and $\varepsilon\, unif(B - B_L)$ over $B - B_L$. As $\varepsilon \to 0$, we recover the node survival we're interested in, but as long as $\varepsilon > 0$, we get ergodicity very easily.

We'll use this as a rhetorical point. In practice, once

ε is small enough, we'd never see an input from $B-B_L$ in a ~~finite~~ feasible observation time, so we'll use $\varepsilon = 0$ in the implementation. But hey, this satisfies proportionality.

Finally, for independent censoring, that clause "conditional on exposure and covariates" hints that I need to make another appeal to the stationary distribution.

But now, I'm tired. It's time for bed.

12/5/22 I wasn't able to find my elaboration on what "conditional on exposure and covariates" means, but for the model at hand—let's call it the brancher-annihilator—I have an argument in mind:

If the Markov chain is ergodic, then the independent censoring condition is satisfied for the extinction time. Why? Well:

- The only reason a message would be censored is that the observation period ended. The extinction time is ultimately governed by the chain's stationary distribution, so the censored messages have the same survival prospects as any other message.
  - Admittedly, having an expression explicitly relating the extinction time of a message at node i to the stationary distribution of the brancher-annihilator (BrAn) would make this argument airtight.

With this in hand, we add a task to the queue:
4) Express survival time probabilities as a function of the stationary distribution

And now, we can move on to task 3