11/30/22 The goal of this notebook is to produce an article out of the REU research.

The theoretical approach has gotten me nowhere, so I want to take a statistical approach instead. First, we need to establish some notation.

An adjacency matrix A stands in as a digraph on $[n]$, $n \geq 1$. Entry (i,j) of A is 1 if node j projects an edge to node i, 0 if not. There may be loops.

$[k] = \{1, \dots, k\}$, $k \in \mathbb{N}$

$[k]_0 = \{0, 1, \dots, k\}$, $k \in \mathbb{N}$.

$X_n = [1]_0^n$; this is the state space of the dynamical process to be investigated,

$x_{t+1} = f(Ax_t + b_t)$, x_0 given   # Define f
Include some graphics demonstrating a couple iterations.

so $x_t \in X_n$, for all $t \in \mathbb{N}_0$.

b_t is another stochastic process, the input vector, the message vector, the efferent messages, something like that. It is uniformly distributed over

$B_L = \{ b \in [1]_0^n : \|b\|_1 = L \}$,

where $L \in [n]_0$. Entries of b_t are called messages.

Our desire is to estimate the distribution of survival times for messages injected at node i, for each node $i \in [n]$. Let's denote this variable τ_i.

~~Before doing this, I want to know if τ_i is a Markov chain.~~ Actually, the most important thing is to just get an empirical distribution. If we denote the theoretical pmf of τ_i by p_i, then we are estimating p_i with $\hat{p}_i$. To do this, we need: ~~a message object that:~~

- A way to identify message starting nodes.
- A way to compute the age of a message.
- Knowledge of when the message goes extinct.

Now, instead of having a fixed time horizon, we'll let the simulation keep running until we've collected "enough data" to be confident that, for all $i \in [n]$ and for all $\tau \in \mathbb{N}$, $|\hat{p}_i(\tau) - p_i(\tau)| < \varepsilon$.

This is a fair theoretical question to ask, and to answer it, we need to do some research to see

- What mode of convergence this is
- How many messages injected at a given node have to go extinct to get good convergence, i.e., what published bounds exist?

Once this is implemented, we should compute this empirical distribution for an ensemble of graphs A sampled from the graphon $eA = ER(1/2)$.

We have to play with some math now. ~~We do know that, for~~ a fixed $\tau \in \mathbb{N}$ and total ~~a~~ number of messages

Let M be the total number of messages injected and M_τ be the ~~proporti~~ number of those messages...

$$\frac{\#\{\text{messages injected at } i \text{ which died at age } \tau\}}{\#\{\text{messages injected at } i \text{ which died}\} =: N} \xrightarrow{N \to \infty} p_i(\tau)$$

I'm not certain this is correct, because what about messages that never die?

Here's an idea: if there are messages that never die, then we start by putting all of that mass on the maximum age $+1$. Then, we portion this mass out over max age $+1$ to ∞ so that the mass over max age $+1$ is less than the ~~max~~ mass over max age. I think a geometric distribution over $\{\text{max age} + 1, \dots, \infty\}$ would work.

Take adjacency matrix A,

$$\begin{tikzcd} X \arrow[r, "A\cdot + b_t"] & Y = [n+1]^n \arrow[r, "f(\cdot)"] & Z = X \arrow[ll, bend right] \end{tikzcd}$$

The goal is to find a formula for the stationary probabilities. It doesn't matter if the formula is terrible.

Fix $z \in Z$, and let $[z] = f^{-1}(\{z\})$.

Let p_{xz} be transition probabilities.

p_{xz}

Compute Ax. For all values of $b \in B_L$, compute b

$$\begin{tikzcd} {[1]_0^n} \arrow[r, "A\cdot"] & {[n]_0^n} \arrow[r, "\cdot + b_t"] & {[n+1]_0^n} \arrow[r, "f(\cdot)"] & {[1]_0^n} \arrow[lll, bend right, "t+=1"'] \\ X & W & Y & Z \end{tikzcd}$$

$$[2-(ax+by)](ax+by) \vee 0 = \left(2(ax+by) - (ax+by)^2\right) \vee 0$$

$$(2-x)x \vee 0 = (2x - x^2) \vee 0$$

$$[2 - ((2x-x^2) \vee 0)][(2x-x^2) \vee 0]$$

$$\cancel{2 - (2x-x^2) \vee 0} \qquad 2(2x-x^2) \vee 0 - [(2x-x^2) \vee 0]^2$$

- Remember that the real goal is to ~~compute~~ calculate the distribution of ~~message~~ survival times for a message injected at node i, at time 0, under load L, in network A.
  - Note that this does not care about the initial state, x_{-1}, which is for the best because a message sender is unlikely to know that.
- The first thing I should do is take a statistical approach. So, when you have more free time:
  - [ ] Design a simulation to collect data about survival times
  - [ ] Find a simple probability distribution to model the distributions you see
  - [ ] Extract ~~features from the~~ nodewise and ~~graphwise~~ graph wide features and attempt to predict the parameters of that distribution.

~~We want to have~~ 11/30/22

The goal here is to produce an unbiased estimator $\hat{p}_1(\tau)$ for the distribution of extinction times, $p_1(\tau)$.

Loose with notation for now, ~~we~~ say we want to have injected at least N messages at node 1.

- The time at which this happens is random, but at least N, since only 1 message can be injected at node 1 at a time. Lets call this time $\underline{T^*(N)}$. So: $T^*(N)$ is the time at which N nodes have been injected at node 1.

- Then there's a time horizon T. We need $T \geq T^*(N)$, but we still have to make a choice. Essentially, T is the point at which we stop collecting data about the extinction times.

- Now, ~~if we inject~~ by time T, it's possible that more than N messages have been injected at node 1; call this number $\tilde{N}$. So, the set of all messages which were injected at node 1 by time T can be indexed as $\{m_\alpha\}_{\alpha=1}^{\tilde{N}} =: \mathcal{M}$.

- $\mathcal{M}$ can be partitioned into $\mathcal{M}^\bullet$ and $\mathcal{M}^\circ$, where

  $\mathcal{M}^\bullet$ is the set of messages that were extinct by time T, and

  $\mathcal{M}^\circ$ is the set of messages that were not extinct by time T.

For the elements of $\mathcal{M}^\bullet$, the extinction time is known, but for the elements of $\mathcal{M}^\circ$, we only have a lower bound on the extinction time.

I keep coming back to this topic of survival analysis 12

The survival function S gives the probability that a message m ~~survives goes extinct~~ will be extant past a certain time.

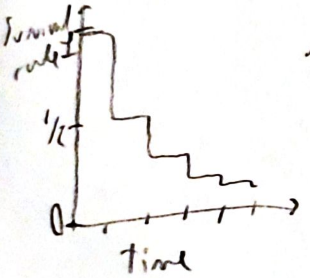
- ext(m) — the extinction age of a message — is a random variable over $\mathbb{N}$. If $a \in \mathbb{N}$ is ~~so~~ the age, then

$S(a) = \mathbb{P}(ext(m) > a)$

Okay. I'll read into the Kaplan-Meier estimator and see if this is the right function for this application.

- Maybe I should change $b_t \overset{unif}{\sim} [1]_0^n$

Kaplan Meier Curve



Survival curve

- How likely is is that message will be extant longer than a certain period of time?

I should skim that "Survival Analysis" book.

1 no 0 yes

Time	Censored
3	1
4	1
4	1
4	0
6	1
7	1
7	1
8	1
9	0
10	1
11	1
13	1
15	0

→

how many deaths at each time

Time	m	q	n
0	0	0	13
3	1	0	13
4	2	1	12
6	1	0	9
7	2	0	8
8	1	1	6
~~9~~	\	\	\
10	1	0	4
11	1	0	3
13	1	1	2
~~15~~	\	\	\

- ☐ Get Kaplan Meier Curve to stand in as empirical distribution
- ☐ Fit different discrete parametric distributions to the KM curve.
- ☐ Regress the parameters on graph properties.