A Directed Web Graph Model with Deletions

Aladdin Final Presentation

Roy Liu

ruoranl@andrew.cmu.edu

Carnegie Mellon University

A Short Web Graph History

- First studied heuristically, starting with Barábasi and Albert in 1999.
- Interest shifted to proving mathematical properties of web graph models (Bollobás et al.; Chung and Lu; Cooper, Frieze, and Vera).

Why Web Graphs?

- Since the growth of web graphs are random, they are nice for modeling large structures like the WWW, whose properties emerge in the limit as time $t \to \infty$.
- The standard random graph model $G(n,p)$ is insufficient for explaining the power law distribution observed on the degree sequence of the WWW. Namely, web graphs have the property that $\mathbf{E}\left[D_k(t)\right]/t \sim Ck^{-\beta}$.
  - $D_k(t)$ is the random variable measuring the number of vertices of degree k at time step t in the web graph
  - C, β positive constants

Overview of the (Simplified) Model

For the sake of simplicity, I will leave out some features of the actual model I proposed. The hardness of the problem is still captured by the features I leave in.

- With probability α, add a vertex to the graph with i_0 in-edges and j_0 out-edges. These are attached preferentially to the rest of the graph based on out-degree and in-degree, respectively.
- With probability $1 - \alpha$, delete a vertex chosen uniformly at random.
- Clean up multi-edges and self-loops.

My Work

- Propose a model combining the ideas of [CFV] and [BBCR].
- Hope that the model gives the joint power law distribution on in-degrees and out-degrees, and prove it. In other words, show that $\mathbf{E}\left[D_{i,j}(t)\right]/t \sim Ci^{-\alpha}j^{-\beta}$.
  - $D_{i,j}(t)$ shall be the random variable measuring the number of vertices of in-degree i, out-degree j at time step t in the web graph
  - C, α, β positive constants

More on Previous Web Graph Results

- [CFV] is an undirected web graph model with deletions of edges and vertices. It involves a three-term, one variable recurrence relation. One can obtain solutions to such a recurrence via Laplace's method.
- [BBCR] is a simple directed web graph model without deletions. It involves solving a two variable recurrence relation. Because of the simplicity of the recurrence, one can obtain solutions via iteration.

The First Attempt

- Modeled after approach used in [CFV]
  - Make simplifying assumptions for recurrence relation on expectations involving error terms and a time variable.
  - Solve homogeneous recurrence first, and then use this result to solve non-homogeneous case.
- Attempted separation of variables (will explain difficulties later).

Recurrence Relations

- Non-homogeneous –

$$\begin{aligned}\mathbf{1}_{i=i_0,j=j_0} = {} & A_0(i-1)f_{i-1,j} + B_0(j-1)f_{i,j-1} \\ & + (A_1 i + B_1 j + E_1)f_{i,j} \\ & + A_2(i+1)f_{i+1,j} + B_2(j+1)f_{i,j+1}\end{aligned}$$

- Homogeneous –

$$\begin{aligned}0 = {} & A_0(i-1)f_{i-1,j} + B_0(j-1)f_{i,j-1} \\ & + (A_1 i + B_1 j + E_1)f_{i,j} \\ & + A_2(i+1)f_{i+1,j} + B_2(j+1)f_{i,j+1}\end{aligned}$$

Separation of Variables

Assume that $f_{i,j} = a_i b_j$. The homogeneous equation "factors" into something like

$$\begin{aligned} 0 = b_j(A_0(i-1)a_{i-1} + (A_1 i + E_1 - k)a_i + A_2(i+1)a_{i+1}) \\ + a_i(B_0(j-1)b_{j-1} + (B_1 j + k)b_j + B_2(j+1)b_{j+1}) \end{aligned}$$

where k is yet to be determined.

Problems with Separation

- What is the k value? What determines it?
- [CFV] had a very clever way of dealing with non-homogeneity. It doesn't work in two-dimensions.
- Other calculations, if they are to be believed, suggest that separation is a bad method.

The Second Attempt

In my second attempt, I used generating functions. Some advantages:

- Can write out the generating function and its associated differential equation fairly easily.
- Generating functions contain all the solutions to a recurrence equation.

Some disadvantages:

- The differential equation in this case was a partial differential equation (more on this).
- Extracting all the solutions involves understanding the analytic behavior of the generating function.

In Series Form

I had trouble integrating, so one portion of the end result expanded in series form was

$$p(x,y) = \frac{(A_0 - A_2)(B_0 - B_2)}{A_0 B_0} \sum_{i,j \geq 1} \frac{1}{(A_0 - A_2)i + (B_0 - B_2)j + E_1} \left(\frac{A_0 x - A_0}{A_0 x - A_2} \right)^i \left(\frac{B_0 y - B_0}{B_0 y - B_2} \right)^j$$

Difficulties in Getting a Solution

- Unable to understand what the coefficients of $p(x, y)$ look like.
- Solving an ODE yields a constant factor which satisfies initial conditions. Solving a PDE yields any differential function on g over ζ, where ζ is some expression in terms of x and y.

The Third Attempt

In my third attempt, I tried the well-studied Polya Urn approach

- Each vertex x_s, where $1 \leq s \leq t$, will have an urn associated with it.
- Each edge incidence on a vertex will be considered a ball in that vertex's urn.
- Balls are thrown in into each urn with probability proportional to the number of balls it contains.

An Example

- Consider two urns with contents i, j. Balls come in one at a time, choosing each urn preferentially. What is the probability that after $k + l - i - j$ balls, the urns have k, l balls, respectively?
- Surprisingly, such a probability is easy to calculate, avoiding consideration of the choice of each ball entirely!

Urns for Web Graphs

- A more complex Polya Urn process is needed to model web graphs, if for each vertex we consider its own urn and a super-urn representing the rest of the graph.
- Consider the urn of some vertex x_s and the super-urn representing the rest of the graph. In addition to throwing in a ball preferentially, we place another ball into the super-urn no matter what (to represent the degree of the newly created vertex). This creates time dependence.

Promise in Urns?

If in some way the time dependence could be dealt with, we could calculate probabilities, which translate into expectation. In other words (assuming the model has no deletions),

$$\sum_{s=1}^{t} \mathbf{Pr}[\deg(x_s) = k] = \mathbf{E}[D_k(t)]$$

In the directed model, this would mean (assuming independence)

$$\sum_{s=1}^{t} \mathbf{Pr}[\deg_{\text{in}}(x_s) = i, \deg_{\text{out}}(x_s) = j]$$

$$= \sum_{s=1}^{t} \mathbf{Pr}[\deg_{\text{in}}(x_s) = i]\mathbf{Pr}[\deg_{\text{out}}(x_s) = j] = \mathbf{E}[D_{i,j}(t)]$$

Current Work

- Would like a close estimate for **Pr**$[\deg(x_s) = k]$, given k, s.
- One can easily derive a power law for the random variable counting the number of vertices of in-degree i, call it $D_i^{\text{in}}(t)$ (or out-degree j). This would mean that

$$\sum_{j \geq 0} D_{i,j}(t) = D_i^{\text{in}}(t) \sim C i^{-\alpha}$$

Unfortunately, summing over the solution obtained from separation of variables causes a disagreement with this new calculation. Explain the inconsistency.

Questions?

References

[BBCR] B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs.

[CFV] C. Cooper, A. Frieze, and J. Vera. Random deletion in a scale free random graph process.