Preferential Attachment Trees with Fitness

Nandan Malhotra

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research, Mohali

April 27, 2021

Certificate of Examination

This is to certify that the dissertation titled "Preferential Attachment Trees with Fitness" submitted by Mr. Nandan Malhotra (Reg. No. MS16133) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Amit Kulshrestha

Dr. Tanusree Khandai

Dr. Neeraja Sahasrabudhe
(Supervisor)

Dated:
27 April 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Neeraja Sahasrabudhe at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgment of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Nandan Malhotra

(Candidate)

Dated:

27 April 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Neeraja Sahasrabudhe

(Supervisor)

Acknowledgment

I would like to acknowledge and express gratitude to my thesis supervisor, Dr. Neeraja Sahasrabudhe (IISER Mohali), for her continuous guidance and support. I would also like to acknowledge and thank my co-supervisor, Dr. Konstantin Avrachenkov (INRIA Sophia Antipolis-Méditerranée), for his valuable collaboration during my thesis. Further, I would like to acknowledge the committee members, Dr. Amit Kulshrestha (IISER Mohali), and Dr. Tanusree Khandai (IISER Mohali), for their valuable feedback, and Dr. Sharayu Moharir (IIT Bombay), for her comments and feedback on the Opinion Dynamics section of this thesis.

I would like to acknowledge the Department of Mathematics of IISER Mohali, for the coursework that helped me during my thesis, INRIA Sophia Antipolis-Méditerranée for the financial support they provided me during my collaboration with the NEO team of INRIA Sophia Antipolis-Méditerranée, and the authors of the papers and texts that I have referred to during this thesis.

I would like to thank my family for their constant love and emotional support, especially during the lockdown period when I had to work remotely from home. Finally, I would like to thank my friends Aishwarya, Anugraha, Kausthub, and Nihkil, for their love, support and aid during my thesis.

Abstract

Preferential Attachment graphs are scale-free growing networks used to model numerous real-world networks. In this thesis, we study degree distributions of directed Preferential Attachment trees with additive fitness. Three regimes of the fitness function are analysed, namely sublinear, linear and superlinear regimes. Further, we obtain analytical expressions for the size of subtree and the height of a vertex in the subtree for a special case of Preferential Attachment with fitness, and use these results to compute expected PageRank for this model. Finally, we study the problem of binary opinion dynamics of a growing population, wherein we obtain the method to determine an optimal influencing strategy to influence the population.

Introduction

Scale free networks are an important class of networks that are used for modelling real world problems. One of the most famous model is the Preferential Attachment Graph, introduced by Barabási and Albert in 1999 and formalized by Bollobás et.al. in 2001. Preferential Attachment graphs exhibit small world property and a power law [1], which make it useful to model real world networks. An important example is citation networks, where each paper can be thought of as a vertex or a node, the outdegree as the number of references, and indegree as the number of citations. Price in 1965 showed that the citation network has a heavy-tailed distribution and follows a power law. In 1976, he proposed a mechanism to explain such an occurrence, which he termed "cumulative advantage", which is now more commonly known as "preferential attachment".

We briefly introduce a model for a Preferential Attachment graph with fitness for three types of fitness [2]. Namely, we study constant fitness, deterministic fitness and random fitness. We focus on directed Preferential Attachment trees with additive fitness, that is, the law of attachment contains an additional fitness parameter apart from the degree of the vertex. The motivation behind this is to incorporate the identity of the vertex into the attachment law. Standard Preferential Attachment laws usually consider the probability of attachment of a newly added vertex to be proportional to the degree of the vertex (such as in [AL06]), or a function of the degree (such as in [DM08], [KRL00], [KR01] and [TGP20]). The paper [AL06] by Avrachenkov and Lebedev provides an analytical expression for the expected PageRank where the law of attachment is proportional to degree. Dereich and Morters in [DM08] explore a model where a new vertex attaches to a random number of nodes with

[1] A network is said to follow a power law when its degree distribution exhibits a power law, that is, the fraction of vertices $P(k)$ of degree k is of the form $P(k) \sim k^{-\gamma}$

[2] Fitness here means a function of the vertex label. This could be deterministic or random.

a sublinear attachment function. They provide an expression for the asymptotic degree distribution for such a model, and a moderate and large deviation principle for temporal evolution of degrees. Krapivsky et.al. in [KRL00] study nonlinear preferential attachment where the function of the degree is of some exponent $\gamma < 1$, and predict the degree sequence to have a power law. In [KR01], Krapivsky and Redner consider $\gamma > 1$ and predict the existence of a single vertex connected to everything else. A similar result was also studied in [OS05] by Oliveira and Spencer, where it was shown that a single dominant vertex emerges and attracts all future edges.

Preferential attachment with fitness preserves the identity of the vertex itself, since the degree of a vertex may not be unique. Fitness is usually classified into two broad categories: additive fitness [3] and multiplicative fitness [4]. Such a model is briefly mentioned in chapter 8, section 8.9 of [vdH14]. The work [GvdHW17] by Garavaglia, Hofstad and Woeginger consider fitness along with aging. By using the theory of aging birth processes, they arrive at an expression for the limiting degree distribution, very similar to that in [DM08]. The work [Ath07] by Athreya considers weighted degrees, which is analogous to multiplicative fitness. He considers a superlinear and sublinear regime for the weight function and considers an embedding in a continuous time pure birth Markov chain to analyse degree sequences. Ergun and Rogers in [ER02] consider a random additive fitness, whereas Bianconi and Barabási consider random multiplicative fitness in [BB01]. Borgs, Chayes, Daskalakis and Roch explore multiplicative fitness further in [BCDR07] and study three phases, namely "first-mover-advantage", "fit-get-richer" and "innovation-pays-off", which helps in determining the "quality" of the vertices for ranking algorithms.

We now give a brief overview of the chapters of this thesis. Chapter 1 gives a brief about Preferential Attachment graphs and gives a power law heuristic for the infamous Barabási-Albert model. We move on to discussing a few extensions of this model and define the model for Preferential attachment with fitness. We then analyse degree distributions for three regimes of the fitness function, namely the sublinear, linear and superlinear regimes. The main tools for this analysis are standard concentration inequalities and concentration

[3] briefly speaking, a function of the vertex is added to a function of the degree

[4] a function of the vertex is multiplied to the function of the degree

of degree sequences.

In chapter 2, we explore the links between urn processes and Preferential Attachment graphs, and illustrate the urns associated with three different types of additive fitness, which are constant fitness, vertex dependent deterministic fitness, and random fitness. We use the theory of Pólya urns to study the size of a subtree rooted at a given vertex for the first case. We then move on to deriving the expression for the probability generating function of the height of a given vertex in a subtree for this case.

In chapter 3, we study an important centrality measure known as PageRank. The motivation behind studying PageRank is the power law hypothesis [5]. We define a Markov chain formulation of PageRank and give necessary conditions that must be fulfilled by a Preferential Attachment law such that the parameter m, which is the number of edges an incoming vertex attaches to the existing graph, does not affect expected PageRank. Using results from chapter 2, we derive an expression for the expected PageRank for the constant fitness model. We also briefly explore another centrality measure known as Closeness Centrality.

In chapter 4, we study influencing strategies for a growing population. We use the stochastic approximation scheme defined by Borkar(2008) to obtain an ODE for our model, and prove a martingale concentration result for the solution for this ODE and a recurrence relation obtained for the model. We then use the ODE solution to determine optimal strategies under certain conditions. Moreover, using simulations, we make certain observations regarding optimality.

The appendix lists out some important results that we have used in the thesis.

Throughout this thesis, we stick to the following notation.

- $\mathcal{O}(f(t))$ denotes the big-O notation, where if $g(t) = \mathcal{O}(f(t))$, then $\exists$ a real $M > 0$ and $t_0 \in \mathbb{R}$ such that

$$|g(t)| \leq Mg(t), \forall t > t_0$$

[5]The power law hypothesis states that in a directed network whose in-degree distribution follows a power law, the PageRank scores will also follow a power law with the same exponent.

- $\mathscr{G}_t$ denotes a graph realization at time t
- $\{\mathscr{F}_t\}_{t\geq 0}$ denotes a filtration
- w.p. is an abbreviation for "with probability"
- $Poi(.)$ denotes the Poisson random variable and $Unif(.)$ denotes the uniform random variable
- $\underline{1}$ denotes a unit vector of ones
- $\mathbf{x}^T$ denotes the transpose of a vector $\mathbf{x}$
- $\mathbb{R}_+^*$ denotes the set of all positive, finite reals

Contents

Chapter 1

Preferential Attachment Graphs

1.1 Introduction

1.1.1 Preferential Attachment models

The study of *Random Graphs* is a mathematically rich area of probability theory, and has numerous applications to other fields. Analysing complex networks is a major motivation behind studying random graphs.

Example 1.1.1 (Erdős-Rényi Random Graph)**.** *For a fixed vertex set V, for any $u, v \in V$, an edge e_{uv} exists with probability p_{uv}. If $p_{uv} = p \ \forall u, v$, then we obtain a homogeneous Erdős-Rényi (ER) Random Graph. ER Random Graphs cannot be used to model real world networks due to their fixed vertex set.*

Most real world networks are characterized by two properties, namely the *small world property* and the *scale free property*. Intuitively, small world property states that distances between vertices is small and thus the graph or network is well connected, and the scale free property states that although the number of vertices with a small degree are common, the number of vertices with a large degree are not too uncommon, that is, the decay for number of vertices of degree at least k is slow for large k. This makes ER random graphs unsuitable to model real world networks, which brings us to another example of random graphs.

Example 1.1.2 (Preferential Attachment Graphs)**.** *$\mathcal{G}_t = (V_t, E_t)$ is a growing random graph where V_t is a growing vertex set. At each time step, one (or more) vertices are added to the graph, and they attach edges to other vertices with probability as a function of their degrees.*

Preferential Attachment Graphs are dynamic in nature and are useful in modelling real world networks. They were shown to have a power law by Yule (1925), and Price (1976) illustrated their scale-free nature.

Definition 1.1.3 (Scale free graphs)**.** *A graph sequence $\{\mathcal{G}_n\}_n$ is said to be sparse if its empirical degree distribution $P_k(n)$ converges to some deterministic limiting probability distribution, that is,*

$$\lim_{n\to\infty} P_k(n) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{d_n(i)=k\}} = p_k$$

where $d_n(i)$ is the degree of vertex i at time n. $\{\mathcal{G}_n\}_n$ is said to be scale-free with exponent τ if it is sparse and

$$\lim_{k\to\infty} \frac{\log p_k}{\log(1/k)} = \tau$$

Barabási and Albert introduced the infamous Barabási-Albert Preferential Attachment model in 1999, which was more rigorously defined and formalized by Bollobás et.al. in 2001, which has been a major model of study in modern research.

Model 1.1.4 (Barabási-Albert Model (Bollobás and Riordan, 2004))**.** *Conditioned on the graph realization $PA_t^{(1,\delta)}$, the law of attachment for an incoming vertex v_{t+1} is given by*

$$\mathbb{P}(v_{t+1}^{(1)} \to v_i^{(1)} | PA_t^{(1,\delta)}) = \begin{cases} \frac{1+\delta}{t(2+\delta)+1+\delta}, & \text{if } i = t+1 \\ \frac{d_t(i)+\delta}{t(2+\delta)+1+\delta}, & \text{if } i \in [t] \end{cases}$$

The sum of degrees of an undirected graph with t vertices where every attaches m edges at every time step is $2mt$. For $m = 1$ as above we get $2t$.

A generalized version of 1.1.4 for m edges can be constructed by identifying vertices $v_{(j-1)m}^{(1)}, \ldots, v_{jm}^{(1)}$ in $PA_{mt}^{(1,\delta/m)}$ and collapsing them into $v_j^{(m)}$ in $PA_t^{(m,\delta)}$. The case where $\delta = 0$ is the standard Barabási-Albert Model. We now illustrate a heuristic for the power law for this model. For simplicity, assume an incoming vertex will not form self loops. Then, we have

$$\mathbb{P}(v_{t+1} \mapsto v_i) = \frac{d_t(i)}{2mt}$$

By assuming the preferential attachment process in continuous time and taking derivative, we obtain

$$\frac{d}{dt} d_t(i) = mp_i = mcd_t(i) = \frac{d_t(i)}{2t}$$

since the rate of change of degree is precisely the expected number of edges added at each time step. Solving the above ODE yields

$$log\left(\frac{d_t(i)}{d_t(i)}\right) = \frac{1}{2}log\left(\frac{t}{i}\right)$$

Since $d_i(i) = m$, and we choose $i \sim Unif(0,t)$, we see that

$$\mathbb{P}(d_t(i) \geq k) = \mathbb{P}\left(i \leq \left(\frac{m}{k}\right)^2 t\right) = \frac{m^2}{k^2}$$

1.1.2 Extensions

We now explore some examples of preferential attachment graphs.
The Barabási-Albert model is an undirected Preferential attachment graph. One can similarly have a directed preferential attachment law given by the following.

Model 1.1.5. *Given the graph realization $\mathscr{G}_t$, the probability that an incoming vertex will attach its k^{th} directed edge (where $1 \leq k \leq m$) to an existing vertex is given by*

$$\mathbb{P}(v_{t+1} \to v_i | \mathscr{G}_t) = \frac{d_t^{in}(i) + d_t^{out}(i)}{2mt}$$

In the above model, the probability of attaching the k^{th} edge does not depend on attaching the previous edge. We can also have a conditional dependency [1] on the attachment of the previous edge.

Consider 1.1.5 as given where m edges are attached independently of one another. The focus of this thesis will be on the model given below.

Model 1.1.6 (Main model of interest)**.** *At $t = 0$, we have the root vertex '0'.*
Given the graph realization $\mathscr{G}_t$, the probability that an incoming vertex will attach its k^{th} directed edge (where $1 \leq k \leq m$) to an existing vertex is given by

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{f(d_t(v)) + g_t(v)}{\sum_{u=0}^{t} (f(d_t(u)) + g_t(u))}$$

where $d_t(v)$ is the indegree of v at time t.

[1] By this, we mean that the probability of attaching the $(k+1)^{\text{th}}$ edge is conditionally dependent on how the k^{th} edge attached to the graph

We will strictly focus on $m = 1$ for our analysis.

The case when $g_t(u) \equiv 0$ and $f(k) = k + m$ has been explored by numerous papers. The paper [AL06] gives an analytical expression for the PageRank of this model and shows that it follows a power law.

For $g_t(v) \equiv 0$, we have three regimes for f, namely sublinear, linear and superlinear f regimes. The sublinear and superlinear cases are referred to as nonlinear preferential attachment. This has been explored in the paper [KRL00] and [KR01]. The superlinear case was also explored in [OS05], and the sublinear case was studied by [DM08].

A power law is observed in the sublinear as well as linear regimes.

A similar model to 1.1.6 is also explored in [TGP20] where degree distributions for various choices of f are obtained.

Another interesting model is the fitness model. One can have deterministic or random fitness, and multiplicative or additive fitness. Random additive fitness has been explored by [ER02], whereas random multiplicative fitness has been explored by [BB01].

We analyse a few fitness models, which are special cases of 1.1.6. The models are as follows.

Model 1.1.7 (Constant Fitness)**.** *At $t = 0$, we have the root vertex '0'.*

Given the graph realization $\mathscr{G}_t$, the probability that an incoming vertex will attach its k^{th} edge (where $1 \leq k \leq m$) to an existing vertex is given by

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{ad_t(v)) + bm}{amt + bm(t+1)}$$

Model 1.1.8 (Deterministic additive fitness)**.** *At $t = 0$, we have the root vertex '0'.*

Given the graph realization $\mathscr{G}_t$, the probability that an incoming vertex will attach its k^{th} edge (where $1 \leq k \leq m$) to an existing vertex is given by

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{ad_t(v) + mg_t(v)}{amt + bm \sum_{u=0}^{t} g_t(v)}$$

Model 1.1.9 (Random additive fitness)**.** *At $t = 0$, we have the root vertex '0'.*

Given the graph realization $\mathscr{G}_t$, the probability that an incoming vertex will attach its k^{th}

edge (where $1 \leq k \leq m$) to an existing vertex is given by

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{ad_t(v) + m\xi_v}{amt + bm \sum_{u=0}^{t} \xi_u}$$

where $\{\xi_i\}_i$ are i.i.d. random variables.

1.2 Degree Distributions

In this section, we derive a relation between the choice of $f(.)$ and the asymptotic degree distribution $P(.)$. We begin with a few concentration results and a claim for the definition of $P(.)$, using which we arrive at a relation between $f(.)$ and $P(.)$ for different regimes of fitness. Finally, we prove our claim for the concentration of $P(.)$.

1.2.1 Concentration Results

The focus will be on model 1.1.6 with $m = 1$. Thus, our law of attachment is given as follows.

Model. *At $t = 0$, the graph $\mathscr{G}_0$ consists of a single vertex labelled '0'.*
At time $t+1$, conditioned on $\mathscr{G}_t$, an incoming vertex attaches a directed edge to one of the existing vertices with the following probability

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{f(d_t(v)) + g_t(v)}{\sum_{u=0}^{t} (f(d_t(u)) + g_t(u))} \tag{1.1}$$

Existing literature focuses on the case when $g_t(u) \equiv 0$ and arrive at a relation between f and P, where P is the limiting indegree distribution of the graph. Explicitly, $P(k)$ is the probability that a uniformly chosen vertex will asymptotically have an indegree k. In [TGP20], various choices of f are obtained by choosing a particular P. In [DM08], the inverse relation is derived, where P is obtained for a given f. The motivation behind introducing the function g is to see the effect of the "identity" of the vertex itself and not just its indegree. This is particularly useful to study preferential attachment with fitness.

We denote by $N_k(t)$ the number of vertices of indegree k at time t. Our first result will prove a concentration for $\mathbb{E}[N_k(t)]$.

Lemma 1.2.1. *Let* $Z_t := \sum_{u=0}^{t} f(d_t(u))$. *Then, if* $\exists\, K < \infty$ *such that* $\forall i \geq 0$,

$$|f(i+1) - f(i)| \leq K$$

then,

$$\mathbb{P}\left(|Z_t - \mathbb{E}Z_t| \geq \sqrt{32K^2 t \ln(t)}\right) = \mathscr{O}\left(\frac{1}{t^4}\right)$$

Proof. We know $Z_t = \sum_{u=0}^{t} f(d_t(u))$. We can split this summation as $Z_t = \sum_j f(j) N_j(t)$, where $N_j(t)$ is the number of vertices of indegree j at time t.

Define $X_k := Z_k - Z_{k-1}$. Then, for $i \neq j$, X_i and X_j are independent random variables, and $\sum_{k=0}^{t} X_k = Z_t$. Now,

$$\begin{aligned} X_k = Z_k - Z_{k-1} &= \sum_{u=0}^{k} f(d_k(u)) - \sum_{u=0}^{k-1} f(d_{k-1}(u)) \\ &= \sum_{u=0}^{k-1} f(d_k(u)) - f(d_{k-1}(u)) + f(0) \quad (\text{since } d_k(k) = 0) \\ &= f(d_{k-1}(l)+1) - f(d_{k-1}(l)) + f(0), \quad \text{for some } l \end{aligned}$$

(since only a single vertex, say l, will have their indegree increased by 1)

Since $|f(i+1) - f(i)| \leq K$, we have

$$-2K \leq X_k \leq 2K$$

Taking $a_k = -2K$, $b_k = 2K$, and by applying the Hoeffding's inequality (see 5.3.1), we have

$$\mathbb{P}(|Z_t - \mathbb{E}Z_t| \geq \delta) \leq 2\exp\left\{-\frac{2\delta^2}{t(4K)^2}\right\}$$

Let $\delta = \sqrt{32K^2 t \ln(t)}$. Then,

$$\mathbb{P}\left(|Z_t - \mathbb{E}Z_t| \geq \sqrt{32K^2 t \ln(t)}\right) \leq 2\exp\left\{-\frac{64K^2 t \ln(t)}{16K^2 t}\right\} = 2\exp\{-4\ln(t)\} = \mathscr{O}\left(\frac{1}{t^4}\right)$$

□

While the above result gives a concentration for $N_k(t)$ where for large t we can approximate $N_k(t) \approx \mathbb{E}N_k(t)$, one can prove a concentration result using Azuma-Hoeffding inequality as well (see 5.3.4). This involves a Doob Martingale argument, where we define $M_n = \mathbb{E}[N_k(t)|\mathscr{G}_n]$. One can see that $M_0 = \mathbb{E}N_k(t)$ and $M_t = N_k(t)$. The result comes from the text [vdH14] by Hofstad, where it is shown that M_n is a martingale with bounded differences, and thus by Azuma-Hoeffding inequality, one can arrive at a bound for $|M_t - M_0|$.

We now require an expression for the asymptotic degree distribution of the preferential attachment graph. Essentially, we want a concentration for the empirical degree distribution defined as follows.

Given $N_k(t)$ as the number of vertices with indegree k, we define the empirical degree distribution $P_k(t)$ as

$$P_k(t) = \frac{N_k(t)}{t+1}$$

where $N_0(0) = 1$. We aim to arrive at a concentration for $P(k)$, which is the probability that a chosen vertex will asymptotically have indegree k. Thus, we wish to see

$$\lim_{t\to\infty} P_k(t) = P(k)$$

We begin by writing a recursion for $\mathbb{E}[N_k(t)]$. For $k = 0$, given the past $\mathscr{G}_t$, we have

$$N_0(t+1) = \begin{cases} N_0(t)+1, \text{ w.p.} & 1 - \frac{\sum_{u=0}^{t}(f(0)+g(u))\mathbb{1}_{\{d_t(u)=0\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} \\ N_0(t), \text{ w.p} & \frac{\sum_{u=0}^{t}(f(0)+g(u))\mathbb{1}_{\{d_t(u)=0\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} \end{cases} \tag{1.2}$$

and for $k > 0$ we have

$$N_k(t+1) = \begin{cases} N_k(t)+1, \text{ w.p.} & \frac{\sum_{u=0}^{t}(f(k-1)+g(u))\mathbb{1}_{\{d_t(u)=k-1\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} \\ N_k(t)-1, \text{ w.p} & \frac{\sum_{u=0}^{t}(f(k)+g(u))\mathbb{1}_{\{d_t(u)=k\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} \\ N_k(t), \text{ w.p } 1 - \frac{\sum_{u=0}^{t}(f(k-1)+g(u))\mathbb{1}_{\{d_t(u)=k-1\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} - \frac{\sum_{u=0}^{t}(f(k)+g(u))\mathbb{1}_{\{d_t(u)=k\}}}{\sum_{u'=0}^{t} f(d_t(u'))+g(u')} & \end{cases} \tag{1.3}$$

Note that the sum $\sum_{u=0}^{t} f(d_t(u))$ can be rewritten as $\sum_j f(j)N_j(t)$. Using lemma 1.2.1, we can replace $\sum_j f(j)N_j(t)$ by $\sum_j f(j)\mathbb{E}N_j(t) + \mathscr{O}\left(\sqrt{t\ln(t)}\right)$ for large t. For now, we consider the following claim to be true.

Claim 1.2.2. *Consider $P(k)$ as defined above. Then,*

$$P(k) = \lim_{t\to\infty} \frac{\mathbb{E}N_k(t)}{t}$$

We will prove this claim after the main results. We now move on to our final concentration result for $\sum_{u=0}^{t} g(u)$ where $g(u) := X_u$ are random variables.
We use Chernoff bound (see 5.3.2) to obtain concentration bounds.

Lemma 1.2.3 (Bounded random variables)**.** *Let $\{X_i\}_{i=1}^{t}$ be random variables such that $\mathbb{P}(X_i \in [a,b]) = 1\ \forall i$. Let $X = \sum_{i=1}^{t} X_i$ and $\mu = \mathbb{E}[X]$. Then, $\forall \delta > 0$,*

$$\mathbb{P}(|X-\mu| \geq \delta\mu) \leq \exp\left\{-\frac{\delta^2\mu^2}{t(b-a)^2}\right\}$$

Remark 1.2.4. *By choice of $\delta = \frac{(b-a)}{\mu}\sqrt{t\ln t}$, we get*

$$\mathbb{P}\left(|X-\mu| \geq (b-a)\sqrt{t\ln t}\right) \leq \exp\{-\ln t\} = \mathcal{O}\left(\frac{1}{t}\right)$$

So, for large t and for $g(u) := X_u$, where $\{X_i\}_i$ are as in the lemma, we have

$$\sum_{u=0}^{t} g(u) = \mu + \mathcal{O}\left(\sqrt{t\ln t}\right) = \sum_{u=0}^{t} \mu_u + \mathcal{O}\left(\sqrt{t\ln t}\right)$$

where $\mu_i = \mathbb{E}X_i$.

Lemma 1.2.5 (Poisson random variables)**.** *Let $X_i \sim Poi(\lambda)\ \forall i$. Then, for $\delta \in (0,\lambda]$, we have*

$$\mathbb{P}(|X_i-\lambda| \geq \delta) \leq 2\exp\left\{-\frac{c\delta^2}{\lambda}\right\}$$

where $c > 0$ is an absolute constant.

Remark 1.2.6. *Since sum of Poisson random variables is also a Poisson random variable, we have $X = \sum_{i=1}^{t} X_i \sim Poi(\sum_{i=1}^{t}\lambda_i)$. Thus, our inequality becomes*

$$\mathbb{P}\left(\left|X - \sum_{i=0}^{t}\lambda_i\right| \geq \delta\right) \leq 2\exp\left\{-\frac{c\delta^2}{\sum_{i=1}^{t}\lambda_i}\right\}$$

Let $\Lambda = \sup_i \lambda_i$. Then,

$$\sum_{i=1}^{t}\lambda_i \leq t\Lambda$$
$$\implies -\frac{1}{\sum_{i=1}^{t}\lambda_i} \leq -\frac{1}{t\Lambda}$$

By taking $\delta = \sqrt{\frac{\Lambda}{c} t \ln t}$*, we get*

$$\mathbb{P}\left(\left|X - \sum_{i=1}^{t} \lambda_i\right| \geq \sqrt{\frac{\Lambda}{c} t \ln t}\right) \leq 2\exp\left\{-\frac{c\left(\sqrt{\frac{\Lambda}{c} t \ln t}\right)^2}{t\Lambda}\right\} = \mathscr{O}\left(\frac{1}{t}\right)$$

and thus we can write $\sum_{u=0}^{t} X_u = \sum_{u=0}^{t} \lambda_u + \mathscr{O}\left(\sqrt{t \ln t}\right)$

Lemmas 1.2.3 and 1.2.5 are exercises from [Ver].

1.2.2 Main Results

We aim to arrive at a relation between $P(.)$ and the choice of f and the limit of the sum $s(t) := \sum_{u=0}^{t} g(u)$. We consider three regimes for $s(t)$ as follows.

1. Sub-linear: $\lim_{t\to\infty} s(t)/t = 0$.
2. Linear: $\lim_{t\to\infty} s(t)/t = c$, for some constant c.
3. Super-linear: $\lim_{t\to\infty} s(t)/t = \infty$ and for some $n \geq 2$, $\lim_{t\to\infty} s(t)/t^n = c$ for some constant c.

Using the concentration result for $N_k(t)$ and taking conditional expectation of the recurrences in 1.2 and 1.3, we obtain

$$\begin{aligned}
\mathbb{E}[N_0(t+1) - N_0(t)|\mathscr{G}_t] &= 1 - \frac{\sum_{u=0}^{t} (f(0) + g(u)) \mathbb{1}_{\{d_t(u)=0\}}}{\sum_{u'=0}^{t} f(d_t(u')) + g(u')} \\
&= 1 - \frac{f(0)N_0(t) + \sum_{u=0}^{t} (g(u)) \mathbb{1}_{\{d_t(u)=0\}}}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t \ln t}\right)}; \\
\mathbb{E}[N_k(t+1) - N_k(t)|\mathscr{G}_t] &= \frac{\sum_{u=0}^{t} (f(k-1) + g(u)) \mathbb{1}_{\{d_t(u)=k-1\}}}{\sum_{u'=0}^{t} f(d_t(u')) + g(u')} - \frac{\sum_{u=0}^{t} (f(k) + g(u)) \mathbb{1}_{\{d_t(u)=k\}}}{\sum_{u'=0}^{t} f(d_t(u')) + g(u')} \\
&= \frac{\sum_{u=0}^{t} g(u)(\mathbb{1}_{\{d_t(u)=k-1\}} - \mathbb{1}_{\{d_t(u)=k\}})}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t \ln t}\right)} \\
&\quad + \frac{f(k-1)N_{k-1}(t) - f(k)N_k(t)}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t \ln t}\right)}
\end{aligned}$$

Taking expectation, we obtain the following recurrence relation.

$$\begin{aligned}
\mathbb{E}[N_0(t+1)] &= \mathbb{E}[N_0(t)]\left(1-\frac{f(0)}{\sum_j f(j)\mathbb{E}N_j(t)+s(t)+\mathscr{O}\left(\sqrt{t\ln t}\right)}\right) \\
&\quad +1-\frac{\sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u)=0)}{\sum_j f(j)\mathbb{E}N_j(t)+s(t)+\mathscr{O}\left(\sqrt{t\ln t}\right)} \\
\mathbb{E}[N_k(t+1)] &= \mathbb{E}[N_k(t)]\left(1-\frac{f(k)}{\sum_j f(j)\mathbb{E}N_j(t)+s(t)+\mathscr{O}\left(\sqrt{t\ln t}\right)}\right) \\
&\quad +\frac{f(k-1)\mathbb{E}[N_{k-1}(t)]+\sum_{u=0}^{t} g(u)(\mathbb{P}(d_t(u)=k-1)-\mathbb{P}(d_t(u)=k))}{\sum_j f(j)\mathbb{E}N_j(t)+s(t)+\mathscr{O}\left(\sqrt{t\ln t}\right)}
\end{aligned} \tag{1.4}$$

We now state and prove our main results for the degree distribution for various regimes of fitness.

Sublinear regime

Theorem 1.2.7. *For the preferential attachment law in 1.1.6 and for $P(.)$ defined in 1.2.2, with the following conditions*

1. $\exists\, K \geq 0$ *such that* $\forall i,\ |f(i+1)-f(i)| < K$
2. $\sum_j f(j)P(j) = \mu$, *where* $\mu \in \mathbb{R}_+^*$
3. $\frac{s(t)}{t} \to 0$ *as* $t \to \infty$

we have the following relations between f and P.

$$f(k) = \frac{\mu}{P(k)}\sum_{i>k} P(i)$$

or equivalently,

$$P(k) = \frac{\mu}{\mu+f(k)}\prod_{i=0}^{k-1}\frac{f(i)}{\mu+f(i)}$$

Proof. Consider the recursion 1.4. We consider the case $k=0$. By taking $a_t = \mathbb{E}N_0(t)$, and

$$b_t = \frac{tf(0)}{\sum_j f(j)\mathbb{E}N_j(t)+s(t)+\mathscr{O}\left(\sqrt{t\ln t}\right)}$$

and

$$c_t = 1 - \frac{\sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u)=0)}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}$$

we get a recurrence of the form

$$a_{t+1} = a_t\left(1 - \frac{b_t}{t}\right) + c_t$$

Thus, we can apply lemma 5.3.5 to obtain $\lim_{t\to\infty} \frac{\mathbb{E}N_0(t)}{t} = \lim_{t\to\infty} \frac{c_t}{1+b_t}$. By our claim, this is precisely $P(0)$. Now,

$$\lim_{t\to\infty} b_t = \frac{f(0)}{\sum_j f(j)P(j)}$$

and

$$\lim_{t\to\infty} c_t = 1$$

since $\sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u)=0) \le s(t)$.
Define $h(k) = \frac{f(k)}{\sum_j f(j)P(j)}$. Thus,

$$P(0) = \frac{1}{1+h(0)}$$

Similarly, for $k > 0$, take $a_t = \mathbb{E}N_k(t)$,

$$b_t = \frac{tf(k)}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}$$

and

$$c_t = \frac{f(k-1)\mathbb{E}[N_{k-1}(t)] + \sum_{u=0}^{t} g(u)(\mathbb{P}(d_t(u)=k-1) - \mathbb{P}(d_t(u)=k))}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}$$

So,

$$\lim_{t\to\infty} b_t = h(k)$$

and

$$\lim_{t\to\infty} c_t = h(k-1)P(k-1)$$

By the same technique, we obtain

$$P(k) = \frac{h(k-1)P(k-1)}{1+h(k)} = P(0)\prod_{i=1}^{k}\frac{h(i-1)}{1+h(i)} = \frac{1}{1+h(k)}\prod_{i=0}^{k-1}\frac{h(i)}{1+h(i)}$$

On the other hand, we can rewrite our last expression as

$$\begin{aligned} P(k) + h(k)P(k) &= h(k-1)P(k-1) \\ \implies h(k)P(k) &= h(k-1)P(k-1) - P(k) \\ &= h(0)P(0) - \sum_{i=1}^{k} P(i) \\ &= 1 - \sum_{i=0}^{k} P(i) \end{aligned}$$

Lastly, we see that by definition of $h(k)$, we have $h(k) = \frac{f(k)}{\mu}$. Using this in the expressions above yields the desired result. □

This result is similar to the ones obtained in [DM08] and [TGP20].

In the following figure, we take a $log - log$ plot of degree distribution $P(k)$ and $(k+1)^2$ for the sublinear case where $g(u) \equiv 0$, that is, the constant fitness case as in 1.1.7. We observe that our analytical expression matches closely to the simulated plot. An error at the tail end is natural due to low sampling at low probability ends and finite size effects [3].

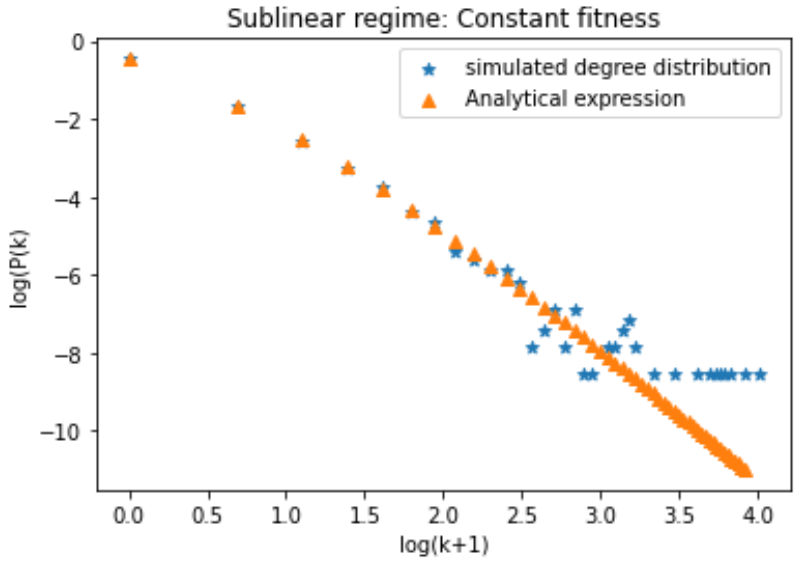


Figure 1.1: A comparison of simulated degree distribution (averaged) and analytical degree distribution for 500 vertices with $f(k) = ak + b$ and $g(u) = 0$

[2]Note that we take $log(k+1)$ instead of $log(k)$, since we will have vertices of degree $k = 0$.

[3]This is a common effect in real world systems, due to which the tail ends tend to fall exponentially rather than in a power-law fashion.

Example 1.2.8. *A few examples of the sublinear cases are as follows.*

1. $g(u) \equiv 0$
2. $g(u) := g_t(u) = \exp(u-t)$
3. $g(u) = \frac{1}{u}$

Example 1.2.9. *Take $f(d_t(u)) = ad_t(u) + b$ and $g(u) \equiv 0$. This is the case when $m = 1$ for 1.1.7. We know that $\mu = \sum_j f(j)P(j)$. Then,*

$$\begin{aligned}
\mu &= \sum_j f(j)P(j) \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_j f(j)\mathbb{E}N_j(t) \quad \textit{(by definition)} \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_j f(j)N_j(t) \quad \textit{(using lemma 1.2.1)} \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} f(d_t(u)) \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} (ad_t(u) + b) \\
&= \lim_{t\to\infty} \frac{at + b(t+1)}{t} = a + b
\end{aligned}$$

Take $a' = a/(a+b)$ and $b' = b/(a+b)$. Taking $F(k) = \frac{f(k)}{\mu}$, we get

$$\begin{aligned}
P(k) &= \frac{1}{1+F(k)} \prod_{i=0}^{k-1} \frac{F(i)}{1+F(i)} \\
&= \frac{\prod_{i=0}^{k-1}(a'i+b')}{\prod_{i=0}^{k}(1+a'i+b')} = \frac{1}{a'} \frac{\prod_{i=0}^{k-1} i + b'/a'}{\prod_{i=0}^{k} i + (b'+1)/a'} \\
&= \frac{1}{a'} \frac{\Gamma(k+\frac{b'}{a'})\Gamma(\frac{1+b'}{a'})}{\Gamma(k+\frac{1+b'+a'}{a'})\Gamma(\frac{b'}{a'})} \xrightarrow[\textit{(using Stirling's Approximation)}]{k\to\infty} \frac{1}{a'} \frac{\Gamma(\frac{1+b'}{a'})}{\Gamma(\frac{b'}{a'})} k^{-\left(1+\frac{1}{a'}\right)} = \frac{1}{\gamma} \frac{\Gamma\left(\frac{1+\beta}{\gamma}\right)}{\Gamma\left(\frac{\beta}{\gamma}\right)} k^{-\left(1+\frac{1}{\gamma}\right)}
\end{aligned}$$

where $\beta, \gamma \in (0,1]$. This is in line with example 1 from [DM08]. One can see that for $\gamma = 1$, we obtain the power law exponent '-2'.

Example 1.2.10. *Consider the case when $g(u) \equiv C$ for some constant C. Although this might not seem like a sublinear case, one can absorb the constant C into the function f and obtain $g(u) \equiv 0$, which is a sublinear case.*

Linear Regime

Theorem 1.2.11. *For the preferential attachment law in 1.1.6 and for $P(.)$ defined in 1.2.2, with the following conditions*

1. $\exists\, K \geq 0$ *such that* $\forall i,\ |f(i+1) - f(i)| < K$
2. $\sum_j f(j)P(j) = \mu$, *where* $\mu \in \mathbb{R}_+^*$
3. $\frac{s(t)}{t} \to 0$ *as* $t \to \infty$

we have the following relations between f and P.

$$f(k) = \frac{\mu + G}{P(k)} \sum_{i>k} P(i) - G$$

or equivalently,

$$P(k) = \frac{\mu + G}{\mu + 2G + f(k)} \prod_{i=0}^{k-1} \frac{f(i) + G}{f(i) + \mu + 2G}$$

Proof. The proof is similar to that of Theorem 1.2.7. For $k = 0$, take $a_t = \mathbb{E}N_0(t)$,

$$b_t = \frac{t f(0)}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathcal{O}\left(\sqrt{t \ln t}\right)}$$

and

$$c_t = 1 - \frac{\sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u) = 0)}{\sum_j f(j)\mathbb{E}N_j(t) + s(t) + \mathcal{O}\left(\sqrt{t \ln t}\right)}$$

Define $h(k) = \frac{f(k)}{\sum_j f(j)P(j) + G}$ and $G_k = \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u) = k)$. We see that

$$\lim_{t\to\infty} b_t = h(0)$$

and

$$\lim_{t\to\infty} c_t = 1 - \frac{G_0}{\mu + G}$$

Thus, $P(0) = \frac{1 - (G_0/(\mu+G))}{1 + h(0)}$.

For $k > 0$, using a similar technique, we obtain

$$\begin{aligned} P(k) &= \frac{h(k-1)P(k-1) + \frac{G_{k-1}-G_k}{\mu+G}}{1+h(k)} \\ \implies h(k)P(k) &= h(k-1)P(k-1) - P(k) + \frac{G_{k-1}-G_k}{\mu+G} \\ &= h(0)P(0) - \sum_{i=1}^{k} P(i) + \frac{G_0 - G_k}{\mu+G}K \\ &= \sum_{i>k} P(i) - \frac{G_k}{\mu+G} \end{aligned}$$

Now, $\sum_k h(k)P(k) + \frac{G}{\mu+G} = \frac{\sum_k f(k)P(k)+G}{\sum_j f(j)P(j)+G} = 1$. Thus,

$$h(k) = \frac{h(k)}{\sum_j h(j)P(j) + G/(\mu+G)} = \frac{(\mu+G)h(k)}{\sum_j (\mu+G)h(j)P(j) + G}$$

Recall that $h(k) = \frac{f(k)}{\sum_j f(j)P(j)+G}$, and we have $f = (\mu+G)h$. Thus,

$$f(k) = (\mu+G)\frac{1}{P(k)}\sum_{i>k} P(i) - \frac{G_k}{P(k)}$$

Finally, we simplify G_k.

We know that $G_k = \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u) = k)$. Recall that $P_k(t)$ is the probability that a chosen vertex at time t has indegree k, and is precisely $\frac{N_k(t)}{t}$. By lemma 1.2.1 and claim 1.2.2, for large t, $P_k(t)$ is close to $P(k)$. In fact, $P_k(t) = P(k) + \mathcal{O}\left(\frac{1}{t} + \sqrt{\frac{\ln t}{t}}\right)$ (which we will see in lemma 1.2.17 in the next section). Then,

$$\begin{aligned} G_k &= \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u) = k) \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} g(u)P_k(t) \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} g(u)(P(k) + \varepsilon_t) \\ &= P(k) \lim_{t\to\infty} \frac{\sum_{u=0}^{t} g(u)}{t} = GP(k) \end{aligned}$$

where ε_t is the error term going to 0. Thus, we get our desired expression for $f(k)$. For the

equivalent expression, we have

$$\begin{aligned} P(k) &= \frac{h(k-1)P(k-1) + \frac{G}{\mu+g}(P(k-1) - P(k))}{1+h(k)} \\ &= \frac{f(k-1)P(k-1) + G(P(k-1) - P(k))}{\mu + G + f(k)} \\ \implies P(k) &= \frac{f(k-1)+G}{f(k)+\mu+2G} P(k-1) \\ &= \frac{\mu+G}{f(k)+\mu+2G} \prod_{i=0}^{k-1} \frac{f(i)+G}{f(i)+\mu+2G} \end{aligned}$$

Similarly,

$$f(k) = (\mu+G)\frac{1}{P(k)} \sum_{i>k} P(i) - \frac{G_k}{P(k)} = \frac{\mu+G}{P(k)} \sum_{i>k} P(i) - G$$

□

Remark 1.2.12. *One can see that for $G = 0$, this reduces to the sublinear case.*

Example 1.2.13. *Consider $g(u) = \alpha + \frac{1}{u}$. Then, since $\sum_{u=0}^{t} \frac{1}{u} \approx \log(t)$, we obtain $G = \alpha$.*

Example 1.2.14 (i.i.d. Poisson random variables)**.** *Let $g(u) := X_u \sim Poi(\lambda)\ \forall u$ and $f(k) = ak + b$. From lemma 1.2.5, we get $G = \lambda$, and we know from example 1.2.9 that $\mu = a + b$. The plot below shows the comparison between analytical and simulated degree distribution.*

In the following figure, we take a $log - log$ plot of degree distribution $P(k)$ and $(k+1)$ for the linear case where $g(u) := X_u \sim Poi(\lambda)$, that is, example 1.2.14. We observe that our analytical expression matches closely to the simulated plot.

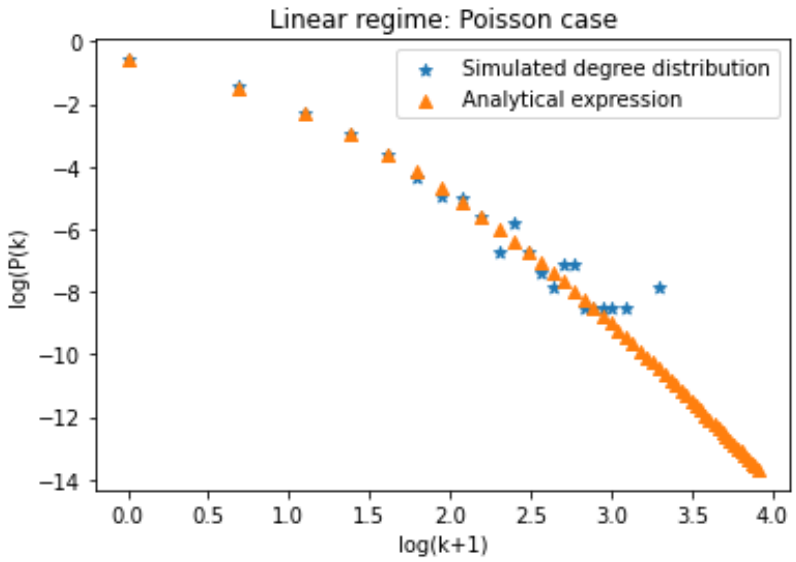


Figure 1.2: A comparison of simulated degree distribution (averaged) and analytical degree distribution for 500 vertices with $f(k) = ak + b$ and $g(u) = Poi(\lambda)$ i.i.d

Superlinear regime

Theorem 1.2.15. *For the preferential attachment law in 1.1.6 and for $P(.)$ defined in 1.2.2, with the following conditions*

1. $\exists\, K \geq 0$ *such that* $\forall i,\ |f(i+1) - f(i)| < K$
2. $\sum_j f(j)P(j) = \mu$, *where* $\mu \in \mathbb{R}_+^*$
3. $\frac{\sum_{u=0}^{t} g(u)}{t} \xrightarrow{t\to\infty} \infty$ *and* $\frac{\sum_{u=0}^{t} g(u)}{t^n} \xrightarrow{t\to\infty} G \neq 0$ *for some* $n \geq 2$.

Then,

$$P(k) = \frac{1}{2^{k+1}}$$

Proof. For $k = 0$, the previous proofs give us

$$b_t = \frac{tf(0)}{\sum_j f(j)\mathbb{E}N_j(t) + \sum_{u=0}^{t} g(u)} \implies \lim_{t\to\infty} b_t = 0$$

and

$$c_t = 1 - \frac{\sum_{u'=0}^{t} g(u')\mathbb{P}(d_t(u') = 0)}{\sum_j f(j)\mathbb{E}N_j(t) + \sum_{u=0}^{t} g(u)} \implies \lim_{t\to\infty} c_t = 1 - \frac{G_0}{G}$$

Thus,

$$P(0) = 1 - \frac{G_0}{G} = 1 - P(0) \implies P(0) = 1/2$$

For $k > 0$, we have from the recursion

$$b_t = \frac{tf(k)}{\sum_j f(j)\mathbb{E}N_j(t) + \sum_{u=0}^{t} g(u)}$$

and

$$c_t = \frac{f(k-1)\mathbb{E}N_{k-1}(t)}{\sum_j f(j)\mathbb{E}N_j(t) \sum_{u=0}^{t} g(u)} + \frac{\sum_{u'=0}^{t} g(u')(\mathbb{P}(d_t(u') = k-1) - \mathbb{P}(d_t(u') = k))}{\sum_j f(j)\mathbb{E}N_j(t) + \sum_{u=0}^{t} g(u)}$$

So, $b = 0$ and $c = \frac{G_{k-1} - G_k}{G}$. By simplifying G_k as before, we get

$$P(k) = \frac{G_{k-1} - G_k}{G} = P(k-1) - P(k) \implies P(k) = P(k-1)/2$$

□

In the following figure, we take a $log - log$ plot of degree distribution $P(k)$ and $(k+1)$ for the superlinear case where $g(u)$ is a polynomial. We observe that the degree of the polynomial does not seem to affect the degree distribution.

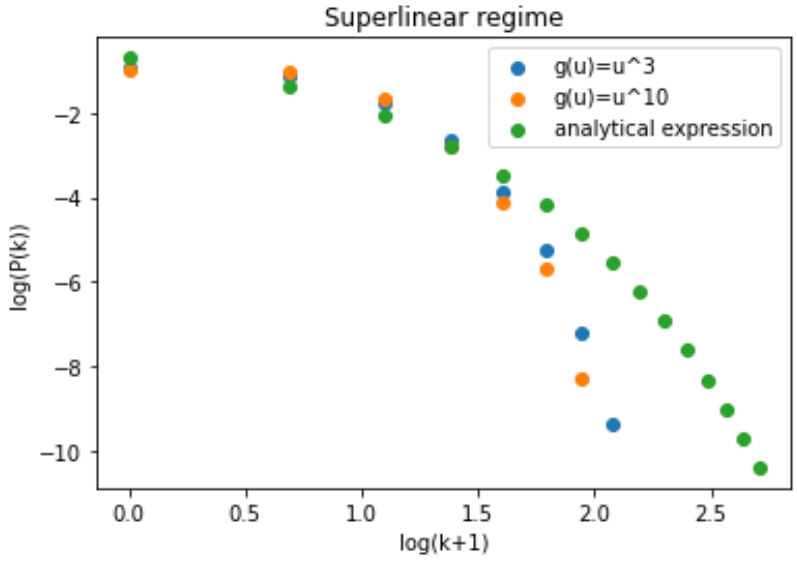


Figure 1.3: A comparison of simulated degree distribution (averaged) and analytical degree distribution for 500 vertices with $f(k) = ak + b$ and $g(u) = u^3$ and $g(u) = u^{10}$

1.2.3 Concentration of degree sequence

We now prove the claim 1.2.2.

From lemma 1.2.1, we already know that $N_k(t) \approx \mathbb{E}N_k(t)$ for large t.

Our proof will be along similar lines as that in [vdH14].

Denote by $\bar{N}_k(t) = \mathbb{E}N_k(t)$.

Proposition 1.2.16. *$\exists$ a constant C such that $\forall t \geq 0$, and all $k \in \mathbb{N} \cup \{0\}$,*

$$|\bar{N}_k(t) - (t+1)P(k)| \leq C$$

Proof. We can write the expressions in 1.4 as

$$\begin{aligned}\bar{N}_k(t+1) = \bar{N}_k(t) + \mathbb{1}_{\{k=0\}} \\ + \frac{f(k-1)\bar{N}_{k-1}(t) - f(k)\bar{N}_k(t) + \sum_{u=0}^{t} g(u)\left(\mathbb{P}(d_t(u)=k-1) - \mathbb{P}(d_t(u)=k)\right)}{\sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}\end{aligned} \tag{1.5}$$

with the convention that $f(-1) = 0$.

If $\bar{N}_k(t) \approx (t+1)P(k)$ for large t, then we have $\bar{N}_k(t+1) - \bar{N}_k(t) \approx P(k)$, and thus the above expression will become

$$P(k) = \mathbb{1}_{\{k=0\}} + \frac{f(k-1)P(k-1) - f(k)P(k) + B_{k-1} - B_k}{\sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}t$$

where $B_i = \frac{\sum_{u=0}^{t} g(u)\mathbb{P}(d_t(u)=i)}{t}$.

Thus,

$$\begin{aligned}(t+2)P(k) &= (t+1)P(k) + P(k) \\ &= (t+1)P(k) + \mathbb{1}_{\{k=0\}} + \frac{f(k-1)P(k-1) - f(k)P(k) + B_{k-1} - B_k}{\sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}t \\ &= (t+1)P(k) + \mathbb{1}_{\{k=0\}} + \frac{f(k-1)P(k-1) - f(k)P(k)}{\sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}t \\ &\quad + \frac{\sum_{u=0}^{t} g(u)\left(\mathbb{P}(d_t(u)=k-1) - \mathbb{P}(d_t(u)=k)\right)}{\sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)}\end{aligned} \tag{1.6}$$

Define $\varepsilon_k(t) = \bar{N}_k(t) - (t+1)P(k)$ and $\mathscr{S} = \sum_j f(j)\bar{N}_j(t) + s(t) + \mathscr{O}\left(\sqrt{t\ln t}\right)$. Then, using 1.5 and 1.6, we have

$$\begin{aligned}
\varepsilon_k(t+1) &= \bar{N}_k(t+1) - (t+2)P(k) \\
&= \varepsilon_k(t) + \frac{f(k-1)}{\mathscr{S}}\left(\varepsilon_k(t-1)\right) - \frac{f(k)}{\mathscr{S}}\left(\varepsilon_k(t)\right) \\
&= \varepsilon_k(t)\left(1 - \frac{f(k)}{\mathscr{S}}\right) + \frac{f(k-1)}{\mathscr{S}}\left(\varepsilon_k(t-1)\right)
\end{aligned}$$

We assume that $\mathscr{S} < \infty$. For $k = 0$, we have

$$|\varepsilon_0(t+1)| = |\varepsilon_0(t)|\left(1 - \frac{f(0)}{\mathscr{S}}\right) < C$$

since f are bounded for lemma 1.2.1 to hold. Similarly, for $k > 0$, we have

$$|\varepsilon_k(t+1)| \leq |\varepsilon_k(t)|\left(1 - \frac{f(k)}{\mathscr{S}}\right) + |\varepsilon_k(t-1)|\frac{f(k-1)}{\mathscr{S}} < C$$

□

We now use lemma 1.2.1 and proposition 1.2.16 as key ingredients in our next lemma.

Lemma 1.2.17. $\mathbb{P}\left(|P_k(t) - P(k)| \geq \frac{C}{t}\left(1 + \sqrt{t\ln t}\right)\right) = \mathscr{O}\left(\frac{1}{t^4}\right)$

Proof. From the previous proposition, we have

$$|\bar{N}_k(t) - (t+1)P(k)| \leq C$$

Thus, using this in lemma 1.2.1, we get

$$\mathbb{P}\left(|N_k(t) - (t+1)P(k)| \geq C\left(1 + \sqrt{t\ln t}\right)\right) = \mathscr{O}\left(\frac{1}{t^4}\right)$$

Since $N_k(t) = (t+1)P_k(t)$, we are done. □

Chapter 2

Subtrees in PA graphs

A directed preferential attachment graph for $m = 1$ is a directed tree. Thus, at each vertex v of the graph, a tree structure can be observed. Two useful quantities can be studied in these tree structures, which are as follows.

- Size of the subtree (denote by $Y_v(t)$), which is the number of vertices in the tree structure excluding the root vertex v.

- Height of a vertex in the tree $T_v(t)$ of the vertex v, which we denote by $X(v,s)$, where s is the local time of the vertex in the tree.

Note that while the subtree size is a global phenomenon, the height of a given vertex in a subtree is a local phenomenon. Both of these quantities will prove to be essential ingredients in studying a centrality measure known as PageRank for a preferential attachment graph. The second quantity will additionally be essential to study another centrality measure known as Closeness Centrality. Both of these will be covered in the next chapter.

2.1 Urn Models

Urn Processes are a useful class of random processes. Typically, one studies one or more urns filled with balls of different colours. The classical problem is where one urn is considered with balls of black and white (or any two) colours, reinforced at every time step. One aims to answer questions such as *"What is the fraction of balls of a particular colour after a large period of time?"* or *"How many times does one pick a particular colour over*

n draws?", and so on. An urn process is usually represented by a **reinforcement matrix**. For the classical problem described above, we have a 2×2 matrix given as follows.

$$\mathscr{R} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

where

- a_{11} = number of black balls added if a black ball is picked
- a_{12} = number of black balls added if a white ball is picked
- a_{21} = number of white balls added if a black ball is picked
- a_{11} = number of white balls added if a white ball is picked

We say an urn is *balanced* if $a_{11} + a_{12} = a_{21} + a_{22}$. The study of urn processes dates back to the post-Renaissance era. Although mentioned in the works of Markov (1905-07) and Ehrenfest and Tatyana (1907), the idea was popularized by Eggenberger and Pólya in 1923. The Pólya-Eggenberger urn model is one of the most widely studied urn process, represented by the following schema.

$$\mathscr{R} = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$$

where $s > 0$. Thus, much like the process of preferential attachment, the Pólya-Eggenberger urn follows the property *"rich get richer"*.

The standard Pólya urn has a constant reinforcement matrix. Variations of this include time dependent or random reinforcements, both of which we will see for our preferential attachment models. For the standard case, the random variable that counts the number of times a ball of a particular colour is drawn converges to a **beta distribution** (see Theorem 5.1.1).

We restrict ourselves to the case where f in 1.1.6 is linear and $m = 1$. In particular, we are interested in models 1.1.7, 1.1.8 and 1.1.9.

2.2 Size of a Subtrees and asymptotic fraction of one colour

2.2.1 Preferential Attachment trees and their corresponding urns

Constant Fitness Model

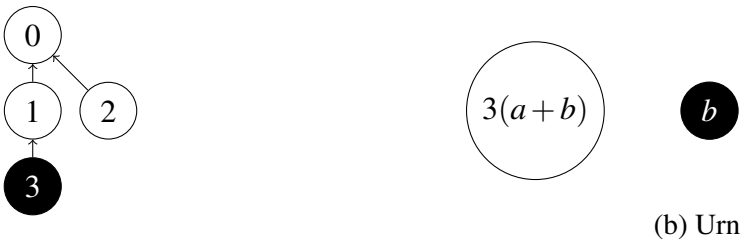
Recall from 1.1.7, the probability of attachment (for $m = 1$) is given by

$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{ad_t(v) + b}{at + b(t+1)}$$

Consider an urn with initial number of black balls $b_0 = v$ and initial number of white balls $w_0 = (a+b)v$, with the reinforcement given by

$$\mathscr{R} = \begin{pmatrix} a+b & 0 \\ 0 & a+b \end{pmatrix}$$

Note that the urn starts at global time $t = v$. Pictorially, we can visualize the urn and the graph as follows (with $v = 3$)

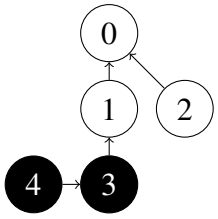


(a) Tree

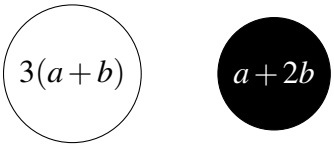
(b) Urn

Figure 2.1: $v = 3, t = 3$

One can see that the probability of attaching to '3' is the same as the probability of picking a black ball, and the probability of attaching to anything but '3' is the same as picking a white ball. Now, assume the incoming vertex "4" attaches to "3",



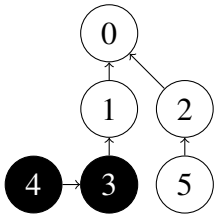
(a) Tree



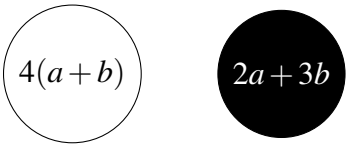
(b) Urn

Figure 2.2: $v = 3, t = 4$

and "5" attaches to "2".



(a) Tree



(b) Urn

Figure 2.3: $v = 3, t = 4$

One can verify that the probability of attaching to the tree structure of "3" (that is either to "3" or "4") is the same as the probability of picking a black ball. Thus, the random variable $Y_v(t)$, which is the number of vertices in the tree structure of v (or the size of $T_v(t)$) is the same as the number of times a black ball is drawn.

Time dependent fitness model

Recall from 1.1.8, the probability of attachment (for $m = 1$) is given by

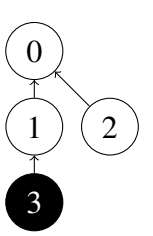
$$\mathbb{P}((t+1) \to v | \mathscr{G}_t) = \frac{ad_t(v) + g(v)}{at + \sum_{u=0}^{t} g(u)}$$

Note that we have not taken any time dependency for g.

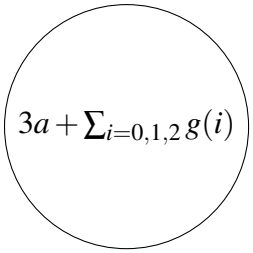
Consider an urn with initial number of black balls $b_0 = g(v)$ and initial number of white balls $w_0 = av + \sum_{i \notin T_v(t)} g(i)$, with the reinforcement given by

$$\mathscr{R}_t = \begin{pmatrix} a+g(t) & 0 \\ 0 & a+g(t) \end{pmatrix}$$

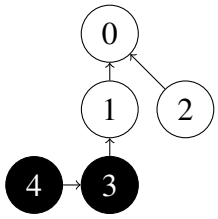
Considering the same graph evolution as in the previous case, for $v = 3$ we have



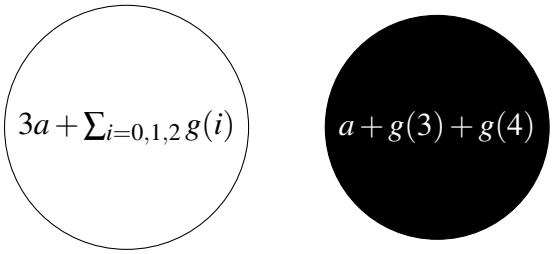
(a) Tree



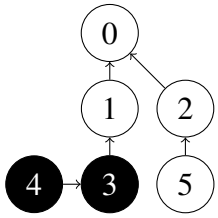
(b) Urn



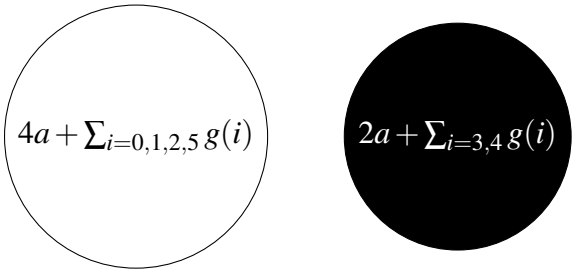
(c) Tree



(d) Urn



(e) Tree



(f) Urn

Figure 2.4: $v = 3$ for $t = 3, 4$ and 5

Thus, we once again see that the random variable $Y_v(t)$ is the same as the number of times a black ball is chosen.

Random Fitness

Recall from 1.1.9, the probability of attachment (for $m = 1$) is given by

$$\mathbb{P}((t+1) \to v|\mathscr{G}_t) = \frac{ad_t(v) + \xi_v}{at + \sum_{u=0}^{t} \xi_u}$$

The relation here is similar as the above case, where we replace $g(u)$ with the random variable ξ_u. Thus, our reinforcement is given by

$$\mathscr{R}_t = \begin{pmatrix} a+\xi_t & 0 \\ 0 & a+\xi_t \end{pmatrix}$$

2.2.2 Main Results

We can now use results for Pólya urns to determine the size of subtree of a preferential attachment graph.

Lemma 2.2.1 (Subtree in a constant fitness model)**.** *For a preferential attachment graph whose law is given by 1.1.7 with $m = 1$, the size of subtree $Y_v(t)$ rooted at v at time time t is given by*

$$\mathbb{P}(Y_v(t) = k) = \frac{\Gamma(t-v+1)\Gamma\left(\frac{b}{a+b}+k\right)\Gamma(t-k)\Gamma\left(\frac{b}{a+b}+v\right)}{\Gamma(t-v-k+1)\Gamma(k+1)\Gamma(v)\Gamma\left(\frac{b}{a+b}\right)\Gamma\left(\frac{b}{a+b}+t\right)}$$

Proof. Using the construction in the previous subsection, we can study a Pólya urn process with a reinforcement $s = a+b$ and $w_0 = (a+b)v$ and $b_0 = b$. Note that the urn evolves for $t-v$ time steps. Then, using theorem 5.1.1

$$\begin{aligned}
\mathbb{P}(Y_v(t) = k) &= \binom{t-v}{k} \frac{b(a+2b)...(b+(k-1)(a+b))}{(b+(a+b)v)(b+(a+b)(v+1))...(b+(t-1)(a+b))} \\
&\times ((a+b)v)((a+b)(v+1))...((t-k-1)(a+b)) \\
&= \binom{t-v}{k} \frac{\left\langle \frac{b}{a+b} \right\rangle_k \left\langle \frac{(a+b)v}{a+b} \right\rangle_{t-v-k}}{\left\langle \frac{b+(a+b)v}{a+b} \right\rangle_{t-v}} \qquad \text{(where } < .. >_k \text{ indicates a prodcut of } k \text{ terms)} \\
&\text{(multiplying and dividing by the appropriate products and using } \Gamma(x) = (x-1)!\text{, we get)} \\
&= \binom{t-v}{k} \frac{\Gamma\left(\frac{b}{a+b}+k\right)\Gamma\left(\frac{(a+b)v}{a+b}+t-v-k\right)\Gamma\left(\frac{b+(a+b)v}{a+b}\right)}{\Gamma\left(\frac{(a+b)v}{a+b}\right)\Gamma\left(\frac{b}{a+b}\right)\Gamma\left(\frac{b+(a+b)v}{a+b}+t-v\right)} \\
&= \frac{\Gamma(t-v+1)\Gamma\left(\frac{b}{a+b}+k\right)\Gamma(t-k)\Gamma\left(\frac{b}{a+b}+v\right)}{\Gamma(t-v-k+1)\Gamma(k+1)\Gamma(v)\Gamma\left(\frac{b}{a+b}\right)\Gamma\left(\frac{b}{a+b}+t\right)}
\end{aligned}$$

□

For $a = b = 1$, the above result coincides with Lemma 5.1 from [AL06].

While for the constant fitness model one can derive an expression for $Y_v(t)$, for the other two cases we get an expression for $B_v(t)$, that is, the number of black balls present in the urn at time t.

Lemma 2.2.2 (Time Dependent fitness)**.** *For a preferential attachment graph whose law is given by 1.1.8 with $m = 1$, with $s(t) = a + f(t)$ where $s(t)$ is a positive integer sequence, then, $\frac{b_n}{\tau_n}$ converges almost surely to a limit $L \in [0,1]$ where $\mathbb{E}L = \frac{b_0}{\tau_0}$, b_0 is number of black balls and τ_0 is total number of balls.*

This result comes from theorem 5.1 from [FM] by Feng and Mahmoud.

Lemma 2.2.3 (Random Fitness)**.** *For a preferential attachment graph whose law is given by 1.1.9 with $m = 1$ and $X := X_t = a + \xi_t$ and $\mu = \mathbb{E}[X]$, b_n converges almost surely to the following*

$$\mu \frac{B}{W+B} t + o(t)$$

where W and B can be characterized as follows:
Let $f_X(s) = s\mathbb{E}(s^X)$ be the probability generating function for $X+1$ and $\phi_B(u) = \mathbb{E}[e^{-uB}]$ and $\phi_W(u) = \mathbb{E}[e^{-uW}]$, then

$$\phi_B^{-1}(v) = (1-v)\exp\left(\int_1^v \left[\frac{\mu}{f_X(s)-s} - \frac{1}{s-1}\right] s\right) = \phi_W^{-1}(v)$$

This result comes from theorem 4 and its subsequent remarks from [Agu09] by Rafik.

2.3 Height of vertex in a subtree

We now find the probability generating function for the height of a vertex in the tree structure of another vertex. We restrict ourselves to the model 1.1.7.

Proposition 2.3.1. *Let $X_s := X(v,s)$ be the height of the s^{th} vertex in the subtree $T_v(n)$ of v. Then,*

$$\mathbb{P}(X_s = k \mid \mathscr{F}_s, Y_v(t) \geq s) = \frac{b\sum_{i=0}^{s-1} \mathbb{1}_{\{X_i = k-1\}} + a\sum_{i=0}^{s-1} \mathbb{1}_{\{X_i = k\}}}{(a+b)s - a}$$

where $\mathscr{F}_s$ is generated by $X_0, X_1, ...X_{s-1}$.

Proof. When studying the random variable X_s, we look at the tree structure of v which contains s vertices, labelled $v_0, v_1, .., v_{s-1}$, where $v_0 = v$. It is trivial to see that $X_0 = 0$ and $X_1 = 1$. One can also see that $X_i = k$ can occur only if $i \geq k$, i.e., one cannot have the event $\{X_s = k\}$ for $k > s$.

For the s^{th} vertex to be a part of the tree structure, its height must lie between 1 and s, that is, it must connect one of the $v_i's$ for $i \in [0, s-1]$. Thus, it has s choices to connect to. The sum of indegrees of these s vertices is $a(s-1)$, and the sum of the "$+b''$ terms from the preferential attachment rule is just bs.

For a fixed k, for X_s to be k, v_s must attach to a vertex of height $k-1$. If there are already vertices of height k, they contribute to the indegree of the vertices of height $k-1$. Thus, for the total number of choices of height $k-1$, given by $\sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k-1\}}$, the sum of indegrees of these choices (which contribute to the preferential attachment law) will be given by $\sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k\}}$. Thus, if K is the scaling factor $at+b(t+1)$ for global time t, we get

$$\begin{aligned}
\mathbb{P}(X_s = k \mid \mathscr{F}_s, Y_v(t) \geq s) &= \mathbb{P}(v_s \mapsto v_i \mid v_i = k-1, i \in [0, s-1]) \\
&= \frac{\sum_{v_i : v_i = k-1, i \in [0,s-1]} (ad_t(v_i) + b)/K}{((a+b)s-a)/K} \\
&= \frac{b \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k-1\}} + a \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k\}}}{(a+b)s-a}
\end{aligned}$$

□

Lemma 2.3.2 (Height of a vertex)**.** *For X_s defined as above,*

$$\mathbb{E}[c^{X_s} \mid Y_v(t) \geq s] = \frac{\Gamma\left(s + \frac{bc}{a+b}\right)\Gamma\left(\frac{b}{a+b}\right)}{\Gamma\left(s + \frac{b}{a+b}\right)\Gamma\left(\frac{bc}{a+b}\right)}$$

Proof.

$$\begin{aligned}
\mathbb{E}[c^{X_s} \mid \mathscr{F}_s, Y_v(t)] &= \frac{\sum_{k=0}^{s} c^k \left[b \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k-1\}} + a \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i=k\}} \right]}{(a+b)s-a} \quad \text{(using the proposition)} \\
&= \frac{a \sum_{k=0}^{s} c^k \mathbb{1}_{\{X_{s-1}=k\}} + b \sum_{k=1}^{s} c^k \mathbb{1}_{\{X_{s-1}=k-1\}}}{(a+b)s-a} \\
&\quad + \frac{((a+b)(s-1)-a)}{(a+b)s-a} \frac{\sum_{k=0}^{s} c^k \left[b \sum_{i=0}^{s-2} \mathbb{1}_{\{X_i=k-1\}} + a \sum_{i=0}^{s-2} \mathbb{1}_{\{X_i=k\}} \right]}{(a+b)(s-1)-a} \\
&= \frac{a \sum_{k=0}^{s} c^k \mathbb{1}_{\{X_{s-1}=k\}} + bc \sum_{k=0}^{s-1} c^k \mathbb{1}_{\{X_{s-1}=k-1\}}}{(a+b)s-a} \quad \text{(change of indices)} \\
&\quad + \frac{((a+b)(s-1)-a)}{(a+b)s-a} \mathbb{E}[c^{X_{s-1}} \mid \mathscr{F}_{s-1}, Y_v(t) \geq s]
\end{aligned}$$

Taking double expectation over $\mathbb{E}[. \mid Y_v(t) \geq s]$, we get

$$\mathbb{E}[c^{X_s} \mid Y_v(t) \geq s] = \frac{\mathbb{E}[c^{X_{s-1}} \mid Y_v(t)]}{(a+b)s-a}(a+bc+(a+b)(s-1)-a)$$

$$\text{(since } \mathbb{1}_{\{X_{s-1}=s\}} = 0, \text{ so } \sum_{k=0}^{s-1} c^k \mathbb{1}_{\{X_{s-1}=k-1\}} = \sum_{k=0}^{s} c^k \mathbb{1}_{\{X_{s-1}=k-1\}})$$

By properties of conditional expectation and using $\mathbb{E}\mathbb{1}_{\{A\}} = \mathbb{P}(A)$, we get

$$\begin{aligned}
\mathbb{E}[c^{X_s} \mid Y_v(t) \geq s] &= \frac{bc+((a+b)s-a)-b}{(a+b)s-a}\mathbb{E}[c^{X_{s-1}} \mid Y_v(t)] \\
&= \left(1 - \frac{b(1-c)}{(a+b)s-a}\right)\mathbb{E}[c^{X_{s-1}} \mid Y_v(t)] \\
&\text{(by using the recursion and } X_0 = 0, \text{ we get)} \\
&= \prod_{k=1}^{s}\left(1 - \frac{b(1-c)}{(a+b)k-a}\right) \\
&= \prod_{k=1}^{s}\left(\frac{(a+b)k-(a+b)+bc}{(a+b)k-a}\right) \\
&= \prod_{k=1}^{s}\left(\frac{k+\frac{bc}{a+b}-1}{k+\frac{b}{a+b}-1}\right) \\
&= \frac{\Gamma\left(s+\frac{bc}{a+b}\right)\Gamma\left(\frac{b}{a+b}\right)}{\Gamma\left(s+\frac{b}{a+b}\right)\Gamma\left(\frac{bc}{a+b}\right)}
\end{aligned}$$

□

For $a = b = 1$, this result coincides with Lemma 6.2 of [AL06].

Chapter 3

Centrality measures for PA graphs

While degree distributions give a reasonable idea of how a graph looks like, they are a very simplistic form of centrality measures used to study graphs. A degree distribution usually only tells us frequencies of degrees in the graph, which can be a helpful indication to point out hubs or existence of concentration of edges onto a certain vertex or vertices. There are other forms of centrality as well, which take into account weights, direction of the edges, number of neighbours and next nearest neighbours and so on. We focus on two popular centrality measures, namely *PageRank* and *Closeness Centrality*.

3.1 PageRank

3.1.1 Introduction

Introduced in 1996 by Larry Page and Sergey Brin, PageRank was introduced as an algorithm for site scoring and page ranking. Motivated by the search engine "RankDex", PageRank laid the foundation for what is now the modern Google search engine. PageRank was influenced by citation networks and Hyper Search, and is a variant of another centrality measure known as EigenCentrality. Like EigenCentrality, PageRank assigns a score based on a node's (or vertex) connections and the connections' connections. Additionally, PageRank also accounts for direction and weights.

The paper [Ye] introduces a Markov chain formulation of PageRank, which is also the definition used in [AL06]. Proofs of results in this section borrow ideas from the latter

paper.

Before formally defining PageRank, we give some preliminaries.

Definition 3.1.1 (Markov Chain)**.** *A (time homogeneous) Markov chain $MC(P)$ is a discrete stochastic process such that*

$$\mathbb{P}(X_{n+1} = x | X_n = x_n, ..., X_1 = x_1) = \mathbb{P}(X_{n+1} = x | X_n = x_n)$$

The transition probabilities p_{ij} of the Markov chain are defined as

$$p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$$

and is independent of time n. The matrix P of entries p_{ij} is called the transition matrix, and the set of all values that the Markov chain can attain is called the state space $\mathscr{S}$.

Definition 3.1.2. *For a Markov chain $MC(P)$, the stationary distribution π is a (row) vector such that*

$$\pi P = \pi;$$
$$\pi \underline{1}^T = 1$$

where $\underline{1}$ is a row vector of ones.

3.1.2 Markov Chain formulation

Definition 3.1.3. *Let P be a transition matrix for n vertices with entries $\{p_{ij}\}_{i,j \in \mathscr{S}}$ (where $\mathscr{S}$ is the state space) where if i has m outgoing links, then $p_{ij} = \frac{k}{m}$ if k links connect i to j. Define $\tilde{P} = cP + \frac{1-c}{n}E$ where E is an $n \times n$ matrix with entries 1. Then, the Pagerank of the network is defined as the stationary distribution π of $\tilde{P}$.*

c is often referred to as the *damping factor* and is a value in (0,1). Google's algorithm chooses $c = 0.85$.

One can see that $\tilde{P}$ is stochastic (since sum of row entries is 1), aperiodic (since due to the matrix E, every state is reachable from another in 1 step), and irreducible (since there always exists a path from one state to another). Thus, there exists a unique vector π that is the stationary distribution of $\tilde{P}$.

Lemma 3.1.4. *The Pagerank of a vertex $v > 0$ at time n is given by*

$$\pi_v(n) = \frac{1-c}{n+1}\left(1 + \sum_{l \in L_v(n)} \left(\frac{c}{m}\right)^{|l|}\right) \tag{3.1}$$

where $L_v(n)$ is the set of all paths that lead to v at time n from vertices $v+1, v+2, ...n$. For the root vertex $v = 0$, we have

$$\pi_0(n) = \frac{1}{n+1}\left(1 + \sum_{l \in L_0(n)} \left(\frac{c}{m}\right)^{|l|}\right) \tag{3.2}$$

Proof. Since the vertex labels start from '0', at time n, the graph consists of $n+1$ vertices. We know that $\tilde{P} = cP + \left(\frac{1+c}{1+n}\right)E$ where $E = [\underline{1}^T\ \underline{1}^T ... \underline{1}^T]$ is a matrix with all entries equal to 1, P is the preferential attachment matrix and I is the identity matrix.

Then,

$$\pi\tilde{P} = \pi cP + \pi\left(\frac{1-c}{1+n}\right)[\underline{1}^T ... \underline{1}^T] = \pi$$

So,

$$\begin{aligned}\pi(I - cP) = \left(\frac{1-c}{1+n}\right)\underline{1} \implies \pi &= \left(\frac{1-c}{1+n}\right)\underline{1}[I - cP]^{-1} \\ &= \left(\frac{1-c}{1+n}\right)(I + \underline{1}\sum_{k\geq 1}(cP)^k)\end{aligned}$$

For some vertex v at time n,

$$\pi_v(n) = \left(\frac{1-c}{1+n}\right)\left(1 + \left(\underline{1}\sum_{k\geq 1}(cP)^k\right)_v\right)$$

where $(..)_v$ denotes the v^{th} entry of the vector.

So, we need to show $(\underline{1}\sum_{k\geq 1}(cP)^k)_v = \sum_{l \in L_v(n)}\left(\frac{c}{m}\right)^{|l|}$

Consider a Markov chain $MC(P)$. Then, for a fixed k, P^k is the k^{th} step transition matrix. A pre-multiplication by c^k ensures that a transition of states occurs due to P and not uniformly due to $\frac{1}{n+1}E$. So,

$$\underline{1}(cP)^k_v = c^k \sum_{i \in \mathscr{S}} p^k_{iv}$$

We know that $p_{iv} = \frac{r}{m} = r\left(\frac{1}{m}\right)$ if r edges connect i to v. Thus, if there are r_i paths of length k connecting i to v, $p^k_{ij} = r_i\left(\frac{1}{m}\right)^k$. Thus,

$$\sum_{i \in \mathscr{S}} p^k_{iv} = \sum_{i \in \mathscr{S}} r_i\left(\frac{1}{m}\right)^k = \sum_{l \in L_v(n)}\left(\frac{1}{m}\right)^{|l|}\mathbb{1}_{\{|l|=k\}}$$

Therefore,

$$\begin{aligned}\pi_v(n) &= \left(\frac{1-c}{1+n}\right)\left(1+\sum_{k\geq 1}\sum_{i\in\mathscr{S}} c^k p_{iv}^k\right)\\ &= \left(\frac{1-c}{1+n}\right)\left(1+\sum_{k\geq 1}\sum_{l\in L_v(n)}\left(\frac{c}{m}\right)^{|l|}\mathbb{1}_{\{|l|=k\}}\right)\\ &= \frac{1-c}{1+n}\left(1+\sum_{l\in L_v(n)}\left(\frac{c}{m}\right)^{|l|}\right)\end{aligned}$$

For $v=0$, $\pi_0(n)=\frac{1}{1+n}\left(1+\sum_{l\in L_0(n)}\left(\frac{c}{m}\right)^{|l|}\right)$, since due to self-loops at 0, every term is multiplied by the series $(1+c+c^2+...)=\frac{1}{1-c}$, which cancels out the $(1-c)$ in the numerator. □

Corollary 3.1.4.1. *For $m=1$, let $T_v(n)=\{w \mid w\mapsto v\}$ be the tree rooted at v. For $w\in T_v(n)$, define $X_n(v,w)$ as the height of w in $T_v(n)$, i.e., the length of the path from w to v.*
Let $Y_v(n)=\mid T_v(n)\mid -1$.
Consider elements of $T_v(n)$ labeled in their local time form, with $v:=$ "0" in the local time of $T_v(n)$. Then

$$\pi_v(n)=\left(\frac{1-c}{1+n}\right)\left(1+\sum_{s=1}^{Y_v(n)} c^{X(v,s)}\right) \tag{3.3}$$

For the root vertex, we have

$$\pi_0(n)=\left(\frac{1}{1+n}\right)\left(1+\sum_{s=1}^{n} c^{X(v,s)}\right) \tag{3.4}$$

We now have a neat expression for PageRank for the case $m=1$. Our interest is now in models for which a general m case can be reduced to $m=1$. This brings us to our main results.

3.1.3 Main Results

Theorem 3.1.5. *Consider a growing preferential attachment graph $\{\mathscr{G}_t\}_t$ with the probability of attachment $\mathscr{P}_t^u:=\mathbb{P}((t+1)\mapsto u\mid\mathscr{G}_t)$ such that*

1. *$\mathscr{P}_v^v=\mathbb{P}((v+1)\mapsto v\mid\mathscr{G}_v)$ is independent of m*
2. *$\mathbb{E}_{\mathscr{G}_t}\mathscr{P}_t^u$ is independent of m, $\forall u$ at time t*

Then, $\mathbb{E}\pi_v^m(n)=\mathbb{E}\pi_v(n)$, that is, m does not affect expected PageRank.

Proof. Let v be a vertex of the growing network. We shall prove this by induction.

At time $t = v+1$, let the incoming vertex $v+1$ attach $k \leq m$ edges to v, and the rest to other vertices. Recall that

$$\pi_v(n) = \frac{1-c}{1+n}\left(1 + \sum_{l \in L_v(n)} \left(\frac{c}{m}\right)^{|l|}\right)$$

So, taking $n = v+1$ and taking expectation, for m edges, we get

$$\begin{aligned}
\mathbb{E}\pi_v(v+1) &= \mathbb{E}\left[\frac{1-c}{1+v+1}\left(1 + \sum_{l \in L_v(v+1)} \left(\frac{c}{m}\right)^{|l|}\right)\right] \\
&= \frac{1-c}{v+2} + \frac{1-c}{v+2}\mathbb{E}\left[\sum_{l \in L_v(v+1)} \left(\frac{c}{m}\right)\right] \qquad \text{(since } |\, l \,| = 1\text{)} \\
&= \frac{1-c}{v+2} + \frac{c(1-c)}{m(v+2)} \sum_{k=1}^{m} k\mathbb{P}(\{\ v+1 \text{ attaches } k \text{ edges to } v\}\ )
\end{aligned}$$

since the expectation term is computing the expected number of paths (of length 1) that go from $v+1$ to v, which is precisely the summation in the next step. Since the probability of attaching k edges from $v+1$ to v is distributed binomially as $Bin(m, \mathscr{P}_v^v)$ the summation term is exactly $m\mathscr{P}_v^v$, and thus

$$\begin{aligned}
\mathbb{E}\pi_v(v+1) &= \frac{1-c}{v+2} + \frac{c(1-c)}{m(v+2)} m\mathscr{P}_v^v \\
&= \frac{1-c}{v+2} + \frac{c(1-c)}{(v+2)}\mathscr{P}_v^v
\end{aligned} \tag{3.5}$$

which is independent of m. At time $t = v+n$, consider the same vertex v. We will now assume that $\forall\ k$ such that $v < k \leq v+n$, $\mathbb{E}\pi_k(v+n)$ is independent of m.

For a given graph realization $\mathscr{G}$ at time $v+n-1$, one can write the pagerank as

$$\begin{aligned}
\pi_v^{\mathscr{G}}(v+n) &= \frac{1-c}{v+n+1}\left(1 + \sum_{l \in L_v(v+n)} \left(\frac{c}{m}\right)^{|l|}\right) = \frac{1-c}{v+n+1}\left(1 + \frac{c}{m}\sum_{l \in L_v(v+n)} \left(\frac{c}{m}\right)^{|l|-1}\right) \\
&= \frac{1-c}{v+n+1}\left(1 + \sum_{k=v+1}^{v+n} \frac{c}{m} \sum_{l' \in L_k(v+n)} \left(\frac{c}{m}\right)^{|l'|} |\, e_{k\to v} \,|\right)
\end{aligned}$$

where $|\, e_{k\to v} \,|$ are the number of edges from k to v. Given to us the graph $\mathscr{G}$ of time $v+n-1$, we know exactly the values of $|\, e_{k\to v} \,|$. Consider some vertex k between $v+1$ and $v+n$, since other vertices will not connect to v. The inner summation computers the paths that reach k, whereas multiplication of this by $|\, e_{k\to v} \,|$ gives us all paths to v. Note that the length of the path till k will be $|l| - 1 = |l'|$, where $l \in L_v(v+n)$ and $l' \in L_k(v+n)$. This is taken care of by the $\frac{c}{m}$ term outside the inner summation.

By opening the brackets and taking the $\frac{1-c}{v+n+1}$ term into the summation, we get

$$\pi_v^{\mathscr{G}}(v+n) = \frac{1-c}{v+n+1} + \sum_{k=v+1}^{v+n} \frac{c}{m}\pi_k^{\mathscr{G}}(v+n) \mid e_{k\to v} \mid$$

$\pi_k^{\mathscr{G}}(v+n)$ depends on vertices after k, whereas $\mid e_{k\to v} \mid$ depends on a vertex before k. Given information of k, and since this growing network is Markovian, the past and future are independent. Thus, by taking expectation over all graph realizations $\mathscr{G}$, for $t = v+n$, we obtain

$$\begin{aligned}\mathbb{E}\pi_v &= \frac{1-c}{v+n+1} + \frac{c}{m}\sum_{k=v+1}^{v+n} \mathbb{E}\pi_k \mathbb{E}[\mid e_{k\to v} \mid] \\ &= \frac{1-c}{v+n+1} + \frac{c}{m}\sum_{k=v+1}^{v+n} \mathbb{E}\pi_k (m\mathbb{P}(B_k))\end{aligned}$$

Since $\mid e_{k\to v} \mid$ is distributed binomially. Here, $\mathbb{P}(B_k) = \mathbb{P}(\{\text{one edge from } k \text{ to } v\})$. Let $A = \frac{1-c}{v+n+1}$, then

$$\begin{aligned}\mathbb{E}\pi_v &= A + \frac{c}{m}\sum_{k=v+1}^{v+n-1} \mathbb{E}\pi_k (m\mathbb{P}(B_k)) + \frac{c}{m}(m\mathbb{P}(B_{v+n})\pi_{v+n}) \\ &= A + c\sum_{k=v+1}^{v+n-1} \mathbb{E}\pi_k \mathbb{P}(B_k) + Ac\mathbb{P}(B_{v+n}) \quad (\text{since } \pi_{v+n}(v+n) = A)\end{aligned}$$

Now, we need to show $\mathbb{P}(B_k)$ is independent of m for all k, since π_k is independent of m by the induction hypothesis. We have

$$\begin{aligned}\mathbb{P}(B_k) = \mathbb{P}(k \mapsto v) &= \sum_{\mathscr{G}} \mathbb{P}(k \mapsto v \mid \mathscr{G}_{k-1} = \mathscr{G})\mathbb{P}(\mathscr{G}_{k-1} = \mathscr{G}) \\ &= \mathbb{E}[\mathbb{P}(k \mapsto v \mid \mathscr{G}_{k-1})] = \mathbb{E}_{\mathscr{G}_{k-1}} \mathscr{P}_k^v\end{aligned} \tag{3.6}$$

Thus,

$$\mathbb{E}\pi_v = A + c\sum_{k=v+1}^{v+n-1} \mathbb{E}\pi_k \mathbb{P}(B_k) + Ac\mathbb{P}(B_{v+n})$$

is independent of m. This completes the proof. □

Example 3.1.6. *Let us consider the model defined by the law in 1.1.7.*

1. $\mathscr{P}_v^v = \frac{b}{at+b(t+1)}$ *(since $d_v(v) = 0$) is independent of m*

2. *For some u and t,*

$$\begin{aligned}\mathbb{E}[d_{t+1}(u) \mid d_t(u)] &= d_t(u) + \mathbb{E}[d_{t+1}(u) - d_t(u) \mid d_t(u)] \\ &= d_t(u) + \mathbb{E}X_e^t \text{ (where } X_e^t \sim Bin(m, \mathscr{P}_t^u) \text{ is the no. of edges attaching to } u) \\ &= d_t(u) + m\mathscr{P}_t^u\end{aligned}$$

By properties of conditional expectation, we have

$$\mathbb{E}[ad_{t+1}(u)+bm] = \mathbb{E}[ad_t(u)+bm] + \frac{m}{m(at+bt+b)}\mathbb{E}[ad_t(u)+bm]$$

Since the above yields a telescopic sum, and $\mathbb{E}[ad_u(u)+bm] = bm$, we see that

$$\mathbb{E}[ad_{t+1}(u)+km] = m\zeta_{t+1}^u$$

Thus,

$$\mathbb{E}_{\mathscr{G}_t}\mathscr{P}_t^u = \frac{\mathbb{E}[ad_t(u)+bm]}{m(at+bt+b)} = \frac{\zeta_t^u}{at+bt+b}$$

which is independent of m.

So, for our model, the expected PageRank is independent of m.

We now give the expression for the expected PageRank of the model defined in 1.1.7.

Theorem 3.1.7 (Expected PageRank)**.** *For the Preferential Attachment graph with law of attachment as in 1.1.7, the expected PageRank is given by*

$$\mathbb{E}\pi_v(n) = \frac{1-c}{1+n}\left(1+\sum_{i=1}^{n-v}\frac{\Gamma(n-v+1)\Gamma(\frac{b}{a+b}+i)\Gamma(n-i)\Gamma(\frac{b}{a+b}+v)}{\Gamma(n-v-i+1)\Gamma(i+1)\Gamma(v)\Gamma(\frac{b}{a+b}+n)}\sum_{s=1}^{i}\left(\frac{\Gamma\left(s+\frac{bc}{a+b}\right)}{\Gamma\left(s+\frac{b}{a+b}\right)\Gamma\left(\frac{bc}{a+b}\right)}\right)\right) \tag{3.7}$$

Proof. The case $m > 1$ can be reduced to $m = 1$ for the preferential attachment model as seen in example 3.1.6. For $m = 1$ by 3.1.4.1, we have

$$\begin{aligned}
\mathbb{E}\pi_v(n) &= \frac{1-c}{1+n}\left(1+\mathbb{E}\sum_{s=1}^{Y_v(n)} c^{X_s}\right) \\
&= \frac{1-c}{1+n}\left(1+\mathbb{E}\sum_{s=1}^{n-v} c^{X_s}\mathbb{1}_{\{Y_v(n)\geq s\}}\right) \\
&\qquad \text{(since the tree } T_v(n) \text{ would have evolved for } n-v \text{ time steps)} \\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\mathbb{E}[c^{X_s}\mathbb{1}_{\{Y_v(n)\geq s\}}]\right) \\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\mathbb{E}[c^{X_s}, \mathbb{1}_{\{Y_v(n)\geq s\}}=1]\right) \\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\sum_{k} k\mathbb{P}(c^{X_s}=k, \mathbb{1}_{\{Y_v(n)\geq s\}}=1)\right)
\end{aligned}$$

By properties of conditional probability, we get

$$\begin{aligned}
\mathbb{E}\pi_v(n) &= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\sum_{k} k\mathbb{P}(c^{X_s}=k \mid \mathbb{1}_{\{Y_v(n)\geq s\}}=1)\mathbb{P}(\mathbb{1}_{\{Y_v(n)\geq s\}}=1)\right)\\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\mathbb{E}[c^{X_s} \mid Y_v(n)\geq s]\mathbb{P}[Y_v(n)\geq s]\right)\\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\mathbb{E}[c^{X_s} \mid Y_v(n)\geq s]\sum_{i=s}^{n-v}\mathbb{P}[Y_v(n)=i]\right)\\
&= \frac{1-c}{1+n}\left(1+\sum_{s=1}^{n-v}\mathbb{E}[c^{X_s} \mid Y_v(n)\geq s]\sum_{i=s}^{n-v}\mathbb{P}[Y_v(n)=i]\right)\\
&= \frac{1-c}{1+n}\left(1+\sum_{i=1}^{n-v}\mathbb{P}[Y_v(n)=i]\sum_{s=1}^{i}\mathbb{E}[c^{X_s} \mid Y_v(n)\geq s]\right)
\end{aligned}$$

By substituting expressions for $\mathbb{P}(Y_v(n)=i)$ and $\mathbb{E}[c^{X_s} \mid Y_v(n)\geq s]$ from Lemmas 2.2.1 and 2.3.2 respectively from the previous chapter, we obtain

$$\begin{aligned}
&\mathbb{E}\pi_v(n)\\
&= \frac{1-c}{1+n}\left(1+\sum_{i=1}^{n-v}\frac{\Gamma(n-v+1)\Gamma(\frac{b}{a+b}+i)\Gamma(n-i)\Gamma(\frac{b}{a+b}+v)}{\Gamma(n-v-i+1)\Gamma(i+1)\Gamma(v)\Gamma(\frac{b}{a+b})\Gamma(\frac{b}{a+b}+n)}\sum_{s=1}^{i}\left(\frac{\Gamma\left(s+\frac{bc}{a+b}\right)\Gamma\left(\frac{b}{a+b}\right)}{\Gamma\left(s+\frac{b}{a+b}\right)\Gamma\left(\frac{bc}{a+b}\right)}\right)\right)\\
&= \frac{1-c}{1+n}\left(1+\sum_{i=1}^{n-v}\frac{\Gamma(n-v+1)\Gamma(\frac{b}{a+b}+i)\Gamma(n-i)\Gamma(\frac{b}{a+b}+v)}{\Gamma(n-v-i+1)\Gamma(i+1)\Gamma(v)\Gamma(\frac{b}{a+b}+n)}\sum_{s=1}^{i}\left(\frac{\Gamma\left(s+\frac{bc}{a+b}\right)}{\Gamma\left(s+\frac{b}{a+b}\right)\Gamma\left(\frac{bc}{a+b}\right)}\right)\right)
\end{aligned}$$

□

The work of Konstantin Avrachenkov and Dmitri Lebedev in [AL06] goes further to simplify the above expression for $a=b=1$ using EKHAD package for Maple. This further yields an expression for the asymptotic distribution of π_v using mean-field approximations. In the following figure, we compare our analytical expression for expected PageRank to a simulated PageRank averaged over multiple iterations.

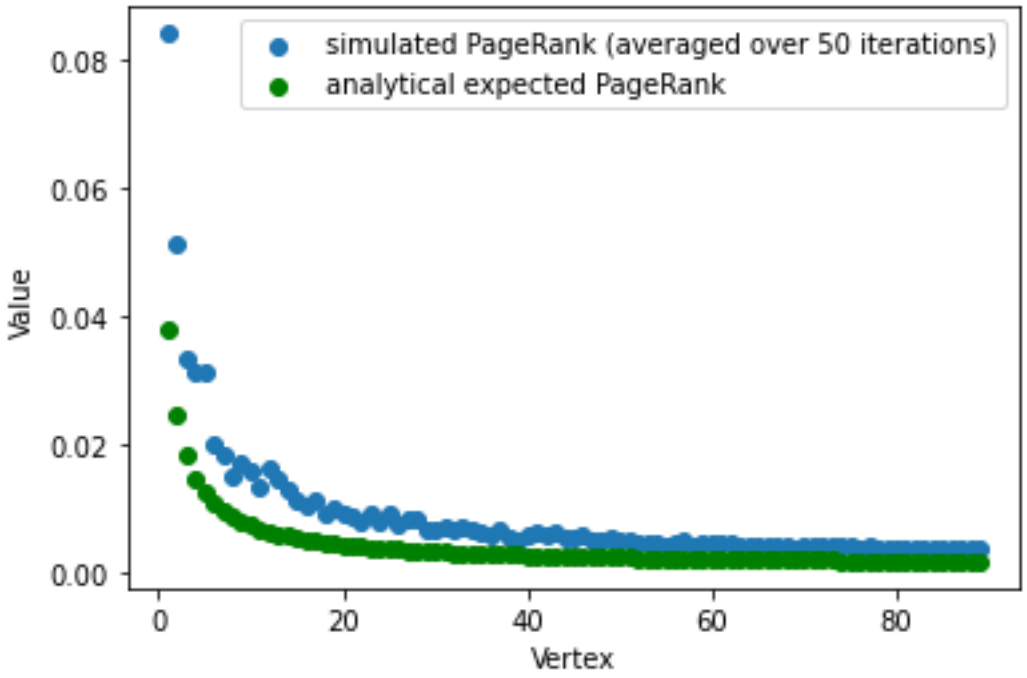


Figure 3.1: A comparison of simulated and analytical expressions for expected PageRank for 90 vertices for the attachment function '$f(k) = 2k + 3$'. Since the analytical expression involves gamma functions, more computational power and memory is required to simulate for a larger number of vertices.

3.2 Closeness Centrality

Another centrality measure is the closeness centrality, which is the reciprocal of "farness". Introduced by Bavelas in 1950, the centrality measure was defined as

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

where $d(y,x)$ is the distance from vertex y to x. For a graph with multiple components, the convention $1/\infty = 0$ is used and the centrality measure is defined as

$$C(x) = \sum_{y \neq x} \frac{1}{d(y,x)}$$

Unlike PageRank, closeness centrality focuses on shortest paths reaching a vertex and assigns a score based on these paths.

Recall from chapter 2 that we have an expression for $\mathbb{P}(X(v,s) = k | \mathscr{F}_s, Y_v(t))$. Let $\chi(v,t)$ be the height of the t^{th} global vertex in the tree of v. Then, the event

$$\{\chi(v,t) = k | t \in Y_v(t), t = s \text{ locally}\}$$

is the same as

$$\{X(v,s) = k\}$$

Thus,

$$\begin{aligned}\mathbb{P}(\chi(v,t) = k, t = s, t \in T_v(t) | \mathscr{F}_s, Y_v(t)) &= \mathbb{P}(\chi(v,t) = k | t = s, t \in T_v(t), \mathscr{F}_s, Y_v(t)) \mathbb{P}(t \in T_v(t)) \\ &= \mathbb{P}(\chi(v,t) = k | t \in T_v(t), \mathscr{F}_s, Y_v(t)) \\ &\times \mathbb{P}(\{\text{a black ball is picked from the urn corresponding to } Y_v(t)\}) \\ &= \left(\frac{a \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i = k\}} + b \sum_{i=0}^{s-1} \mathbb{1}_{\{X_i = k-1\}}}{(a+b)s - a} \right) \left(\frac{(a+b)Y_v(t) + b}{(a+b)t + b} \right)\end{aligned}$$

since the number of black balls is is $(a+b)Y_v(t)+b$ and total number of balls is $(a+b)t+b$. Taking expectation over $\mathscr{F}_s$ we get

$$\mathbb{P}(\chi(v,t) = k, t \in T_v(t) | Y_v(t)) = \left(\frac{\sum_{i=0}^{s-1} (a\mathbb{P}(X_i = k) + b\mathbb{P}(X_i = k-1))}{(a+b)s - a} \right) \left(\frac{(a+b)Y_v(t) + b}{(a+b)t + b} \right) \tag{3.8}$$

So,

$$C_t(v) = \sum_u \frac{1}{\chi(v,u)} \implies \mathbb{E}C_t(v) = \sum_u \mathbb{E}\left[\frac{1}{\chi(v,u)} \right]$$

Therefore,

$$\begin{aligned}\mathbb{E}C_t(v) &= \sum_{u=v+1}^{t} \sum_{k \geq 1} \frac{1}{k} \mathbb{P}(\chi(v,u) = k, u \in T_v(t)) \\ &= \sum_{u=v+1}^{t} \sum_{k \geq 1} \sum_{l \geq 1} \frac{1}{k} \mathbb{P}(\chi(v,u) = k, u \in T_v(t) | Y_v(u) = l) \mathbb{P}(Y_v(u) = l)\end{aligned}$$

Equation 3.8 and lemma 2.2.1 can be used to substitute the two probabilities in the above summation.

Chapter 4

Influencing Opinion Dynamics of growing populations

4.1 Background

Opinion dynamics is an area of study in mathematics, physics and sociology. A commonly studied model is the **voter model**, which is a binary opinion dynamics model. The model can be thought of as a graph with vertices as voters, labelled "0" or "1". At random times, a given vertex flips its opinion or label with a locally defined probability.
The model was introduced by Richard A. Holley and Thomas M. Liggett in 1975. Further work was done by Kempe et.al. (2003) and Grtner and Zehmakan (2017). Kumar et.al. in 2018 studied the conformist and strong-willed variations of this model.

The aim is to influence an evolving population to skew their opinion in your favour. Existing literature focuses on *where* to influence. A novel approach is *when* to influence, with certain constraints. [SKG$^+$20] studies dynamics of a fixed size population of mixed individuals and aims to find the optimal influencing strategy under a time budget. [GMS20] introduces a graph structure to this problem.

Our aim is to find the optimal strategy for a growing population.
We study a model of binary opinion evolution in a growing population that captures the following behaviour:

1. Strong-willed individuals (Type *S*)- Type *S*: these individuals are not influenced by

peers and change their opinions independent of anyone else in the population.

2. Conformist (Type C)- these individuals change their opinion based on the majority (local or global) opinion at that given time and tend to adopt the "popular" opinion at that time.

4.2 Preliminaries

We start with a finite population of M_0 number of people at time $t = 0$, such that each individual has a binary opinion ("Yes" or "No") about a certain (fixed) topic of interest. The evolution of opinions happens in discrete time. At each time step, opinion of the system evolves in two steps:

1. A fixed number of individuals, denoted by $N_c(t)$, are chosen uniformly at random and they change their opinion behaving in one of three ways described above.

2. A fixed number of individuals, denoted by $N_a(t)$ and having opinion 1 with probability α_t and 0 with probability $1-\alpha_t$ are added to the population.

In this work, we assume $N_c(t) = N_c, N_a(t) = N_a$ and $\alpha_t = \alpha$ for all $t \geq 0$. At each time t, depending on the absence or presence of the peer-influences, the individuals chosen for opinion evolution update their opinions organically (S-type behaviour) or get influenced (positively or negatively, i.e., C- or R-type behaviour) by the rest of the population. As in [SKG+20], these models are called Hybrid S/C-type and Hybrid S/R-type. The presence or the absence of peer influence at a given time-step is determined by a parameter $\lambda \in [0,1]$. That is, at any given time, the chosen N_c individuals behave strong-willed with probability of updating her opinion organically λ and are influenced by their peers with probability $1-\lambda$.

Define random variables $\{I_t(i)\}_{1\leq i\leq M_t, t\geq 0}$ taking values in $\{0,1\}$, where $I_i(t)$ denotes the opinion of the i^{th} individual at time t. Thus,

$$I_i(t) = \begin{cases} 1 & \text{if the opinion of } i^{th} \text{ individual at time t is } \textit{Yes} \\ 0 & \text{if the opinion of } i^{th} \text{ individual at time t is } \textit{No} \end{cases}$$

Note that $M_{t+1} = M_t + N_a$. Thus, M_t increases linearly and deterministically. Define random variables: $Y_t = \sum_{i=1}^{M_t} I_t(i)$ and $N_t = M_t - Y_t$ as total number of people with opinion "Yes" and the total number of people with opinion "No" at time t respectively.

As in the earlier models, an influencing agency can manipulate the transition probabilities between the two opinions of any given individual in any given time-slot. However, the influencing agency is assumed to have a time-budget constraint. That is, the agency can only influence a fraction bT is the total time slots T, where b is fixed throughout. Thus, at each time-slot a uniformly selected collection of N_c individual change their opinion from "Yes" to "No" or vice-versa depending on their currently held opinion, the presence/absence of the peer influence and the presence/absence of external influence. The aim of the influencing agency is to target the ']lq correct" time-slots so to ensure that the expected number of people holding opinion "Yes" at time T is maximized. We aim to determine the best strategies for this. In other words, we want to determine which slots should be influenced maximize $\mathbb{E}[Y_T]$. An influencing strategy $\mathscr{S}$ consists of the time-slots $(t_{i_1}, \ldots, t_{i_{bT}})$ that should be influenced. If strategy $\mathscr{S}_1$ is better than strategy $\mathscr{S}_2$, it is denoted by $\mathscr{S}_1 \gg \mathscr{S}_2$.

Definition 4.2.1 (Optimal Strategy)**.** *We call a strategy optimal if the influence according to that strategy results in a larger expected number of "Yes" at the end of time T than the expected number of "Yes" at the end of time T using any other influence strategy.*

Thus, an optimal strategy $\mathscr{S}^*$ is such that $\mathscr{S}^* \gg \mathscr{S}$, where $\mathscr{S}$ is any other collection of bT time-slots to be influenced. As we shall see, due to monotonicity, in most cases influencing the first or the last bT slots is optimal. We denote these strategies by $\mathscr{S}_F$ and $\mathscr{S}_L$ respectively.

4.2.1 Model Dynamics

We begin by writing the evolution of $X_t = Y_t/M_t$, that is, the fraction of individuals of opinion 1 at time t. We divide the discussion into two cases: (i) $N_c = k$ for a fixed k in [1, M_0], and (ii) $N_c = M_t$ for all t.

We will write a general recursion for X_{t+1}.

$$\begin{aligned} X_{t+1} &= \tfrac{Y_{t+1}}{M_{t+1}} = \tfrac{Y_t + I_{t+1}}{M_{t+1}} \\ &= \tfrac{M_t}{M_{t+1}} X_t + \tfrac{I_{t+1}}{M_{t+1}} \end{aligned} \tag{4.1}$$

where I_{t+1} is the change in the number of people of opinion 1 from tine t to $t+1$.

- $N_c = k$**, where** $1 \leq k \leq M_0$ **is a constant (**M_0 **is the initial size of the population)**

  The realizations of the random variable I_{t+1} depend on the two independent processes described in the preliminary section, that is, due to change in opinion of the population followed by addition of new individuals. Thus, we can write I_{t+1} as $I_{t+1} = O^{N_c}_{t+1} + O^{N_a}_{t+1}$, where $O^{N_c}_{t+1}$ is the change due to opinion evolution and $O^{N_a}_{t+1}$ is the change due to newly added individuals.

  Thus,

$$\begin{aligned}\mathbb{E}[I_{t+1}|\mathscr{F}_t] &= \mathbb{E}[O^{N_c}_{t+1} + O^{N_a}_{t+1}|\mathscr{F}_t] \\ &= \mathbb{E}[O^{N_c}_{t+1}|\mathscr{F}_t] + \alpha N_a\end{aligned}$$

  Now,

$$\begin{aligned}\mathbb{E}[O^{N_c}_{t+1}|\mathscr{F}_t] &= \mathbb{E}[\sum_{i=1}^{M_t} O_i(t+1)|\mathscr{F}_t] \\ &= \sum_{i=1}^{M_t} \mathbb{E}[O_i(t+1)|\mathscr{F}_t]\end{aligned}$$

  where $O_i(t+1)$ is the opinion change of the i^{th} individual at time $t+1$.

  We see that

$$\mathbb{E}[O_i(t+1)|\mathscr{F}_t] = ((1-X_t)q_t - X_t p_t)\mathbb{P}(i \in C_k)$$

  where C_k is the set of k chosen individuals whose opinion changes, and so $\mathbb{E}[O_i(t+1)|\mathscr{F}_t] = ((1-X_t)q_t - X_t p_t)\frac{k}{M_t}$. Thus, $\mathbb{E}[I_{t+1}|\mathscr{F}_t] = k((1-X_t)q_t - X_t p_t) + \alpha N_a$.

- $N_c = M_t$

  The major difference for this case is that *all* individuals of the population change their opinion. Thus, one does not have to uniformly pick individuals, due to which $\mathbb{E}[O_i(t+1)|\mathscr{F}_t]$ is just $((1-X_t)q_t - X_t p_t)$.

  By a similar procedure as in the above case, we obtain the following recursion.

$$\mathbb{E}[I_{t+1}|\mathscr{F}_t] = M_t((1-X_t)q_t - X_t p_t) + \alpha N_a \tag{4.2}$$

  We will see in the next section why this case needs to be handled differently.

4.2.2 Stochastic Approximation for our model

The classical stochastic approximation scheme is given by the following iteration for $x \in \mathbb{R}^d$

$$x(n+1) = x(n) + a(n)\left[h(x(n)) + \mathscr{M}(n+1)\right],\ n \geq 0, \tag{4.3}$$

such that:

1. $\{a(n)\}$ is a positive step-size sequence satisfying

$$\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty.$$

2. $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz.

3. $\{\mathscr{M}_n\}_{n\geq 0}$ is a square-integrable Martingale difference sequence with respect to a suitable filtration.

4. $\sup_n \|x(n)\| < \infty$.

Then, from the stochastic approximation theory (see 5.2), we know that the iterates of (4.3) converge almost surely to the stable equilibria of the solutions of the O.D.E. asymptotically with probability one.

$$\dot{x}(t) = h(x(t)),\ t \geq 0. \tag{4.4}$$

We now write a stochastic approximation scheme for our model. By adding and subtracting $\mathbb{E}[I_{t+1}|\mathscr{F}_t]$ in , we get

$$\begin{aligned} X_{t+1} &= \frac{M_{t+1} - N_a}{M_{t+1}} X_t + \frac{\mathbb{E}[I_{t+1}|\mathscr{F}_t]}{M_{t+1}} + \frac{I_{t+1} - \mathbb{E}[I_{t+1}|\mathscr{F}_t]}{M_{t+1}} \\ &= X_t + \frac{1}{M_{t+1}}\left[\mathbb{E}[I_{t+1}|\mathscr{F}_t] - N_a X_t\right] + \frac{\mu_t}{M_{t+1}} \end{aligned}$$

where $\mu_t = I_{t+1} - \mathbb{E}[I_{t+1}|\mathscr{F}_t]$, and $\mathbb{E}[\mu_t|\mathscr{F}_t] = 0 = \mathbb{E}[\mu_t]$.

- $N_c = k$ for a fixed k

  In this case, we get:

$$X_{t+1} = X_t + \frac{1}{M_{t+1}}[\alpha N_a + kq_t - (kq_t + kp_t + N_a)X_t] + \frac{\mu_t}{M_{t+1}}$$

Substituting $p_t = \lambda p + (1-\lambda)p(1-X_t)$ and $q_t = \lambda q + (1-\lambda)qX_t$, we get

$$X_{t+1} = X_t + \frac{1}{M_{t+1}}[k(\lambda-1)(q-p)X_t^2 - (kp+(2\lambda-1)kq+N_a)X_t + \lambda kq + \alpha N_a] + \frac{\mu_t}{M_{t+1}} \tag{4.5}$$

Note that conditions 4.2.2 are satisfied. In particular,

(i) $M_{t+1} = M_t + N_a = M_0 + (t+1)N_a$. This is a linear step size, and thus $\sum_t M_t = \infty$ and $\sum_t M_t^2 < \infty$.

(ii) Here, $h(x) = k(\lambda-1)(q-p)x^2 - (kp+(2\lambda-1)kq+N_a)x + \lambda kq + \alpha N_a$, which is a quadratic function in x. For $\lambda = 1$ or $p = q$, the function is linear and thus Lipschitz. For other cases, $|h(x_2) - h(x_1)| \leq |(x_2 - x_1)||k(\lambda-1)(q-p)(x_2 + x_1) - (kp+(2\lambda-1)kq+N_a)| = K|x_2 - x_1|$, where $0 < K < \infty$ since $x \leq 1$. Thus, the function is Lipschitz.

(iii) $\{\mu_t\}_t$ is a Martingale difference sequence

(iv) Since X_t is the fraction of individuals with opinion 1, $\sup_t ||x_t|| < \infty$.

Now, $\mathscr{D}(x) := \frac{\partial h}{\partial x} = 2k(\lambda-1)(q-p)x - (kp+(2\lambda-1)kq+N_a)$.

For $\lambda = 1$ or $q = p$, $\mathscr{D}(x) < 0$ for all x.

For other cases, consider the quadratic equation. Let its roots be A_1 and A_2 with $A_1 > A_2$. Note that $A_1A_2 = \frac{\lambda kq + \alpha N_a}{k(\lambda-1)(q-p)}$ and $A_1 + A_2 = \frac{(kp+(2\lambda-1)kq+N_a)}{k(\lambda-1)(q-p)}$. We need to check the stability of the partial derivative at these roots, i.e., we need to see the sign of $\mathscr{D}(x) = k(\lambda-1)(q-p)[2x - (A_1+A_2)]$ at $x = A_1, A_2$.

For the case $q > p$ for a general λ, the product of roots of the quadratic equation is negative, which means $A_1 > 0$ and $A_2 < 0$. We can see that $\mathscr{D}(A_2) > 0$, which is unstable, and $\mathscr{D}(A_1) < 0$, which is stable.

For the case $q < p$ for a general λ, the product of roots is positive, which means both roots are positive, since the fraction cannot be negative. Then, since $A_1 > A_2$, $\mathscr{D}(A_1) > 0$ and $\mathscr{D}(A_2) < 0$, thus A_2 is a stable solution.

Thus, the iterates of the recursion (4.5) converge to the stable solutions of the ODE:

$$\frac{dx_t}{dt} = \frac{1}{M_0 + (t+1)N_a}(\alpha N_a + kq_t - (kq_t + kp_t + N_a)x_t) \tag{4.6}$$

That is, $X_t \xrightarrow{a.s.} \frac{\alpha N_a}{kq+kp+N_a}$, for $q_t = q, p_t = p \; \forall t \geq 0$. Note that when $N_a = k$ the limit only depends on p, q and α.

Substituting $p_t = \lambda p + (1-\lambda)p(1-X_t)$ and $q_t = \lambda q + (1-\lambda)qX_t$ in (4.6), we get

$$\frac{dx_t}{dt} = \frac{k(\lambda-1)(q-p)x_t^2 - (kp + (2\lambda-1)kq + N_a)x_t + \lambda kq + \alpha N_a}{M_0 + (t+1)N_a} \tag{4.7}$$

1. $\lambda = 1$

   We focus on the ODE

   $$\frac{dx_t}{dt} = \frac{-(kq + kp + N_a)x_t + kq + \alpha N_a}{M_0 + (t+1)N_a}$$

   the solution to which is given by

   $$x_t = \frac{kq + \alpha N_a}{kp + kq + N_a} + \left(x_0 - \frac{kq + \alpha N_a}{kp + kq + N_a}\right)\left(\frac{t + 1 + M_0/N_a}{1 + M_0/N_a}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)} \tag{4.8}$$

2. $p = q = \rho$

   $$\frac{dx_t}{dt} = \frac{-(2\rho\lambda k + N_a)x_t + \lambda k\rho + \alpha N_a}{M_0 + (t+1)N_a}$$

   with the solution

   $$x_t = \frac{\lambda k\rho + \alpha N_a}{2\lambda k\rho + N_a} + \left(x_0 - \frac{\lambda k\rho + \alpha N_a}{2\lambda k\rho + N_a}\right)\left(\frac{t + 1 + M_0/N_a}{1 + M_0/N_a}\right)^{-\left(\frac{2\lambda k\rho}{N_a}+1\right)} \tag{4.9}$$

3. General Case

   The solution in this case is quite complicated, and is given by the following expression

   $$\frac{x_t - A_1}{x_t - A_2} = \left(\frac{x_0 - A_1}{x_0 - A_2}\right)\left(\frac{t + 1 + M_0/N_a}{1 + M_0/N_a}\right)^{-\frac{k(1-\lambda)(q-p)(A_1-A_2)}{N_a}}$$

   where $A_1 > A_2$ are the real distinct roots of the quadratic

   $$P(x_t) = k(\lambda-1)(q-p)x_t^2 - (kp + (2\lambda-1)kq + N_a)x_t + \lambda kq + \alpha N_a$$

- $N_c = M_t$

  The differential equation cannot be solved using the variable separable method, since the function $h := h(x_t, t)$ has a time t parameter.

4.3 Main results

4.3.1 Martingale Concentration

Note that while the stochastic approximation theory tells us that the recursion will converge almost surely to the stable fixed point of the corresponding ODE, for large t, the trajectories of the recursion and the solution fo the ODE are "close". We now give a rigorous error bound on the solution of the recursion and the approximate solution of the ODE for the cases: $\lambda = 1$ and $p = q$.

Proposition 4.3.1. *For the recursion* (4.5)*, with* $\lambda = 1, p_t = p, q_t = q$ *(or with* $q_t = p_t = \rho$*), for sufficiently large* T*, given* $\varepsilon > 0$*,*

$$P(|X_T - x_T^*| > \varepsilon) < \mu,$$

where x_T^* *is the solution of the ODE given by* (4.8) *(or* (4.9) *respectively), and where* μ *is given by*

$$\mu = 2\exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{T} c_i^2}\right)$$

where

$$c_t = X_t\left(\prod_{i=0}^{t-1} 1 - \frac{kp+kq+N_a}{M_{i+1}}\right)^{-1} - X_{t-1}\left(\prod_{i=0}^{t-2} 1 - \frac{kp+kq+N_a}{M_{i+1}}\right)^{-1} - \frac{kq+\alpha N_a}{M_t}\left(\prod_{j=0}^{t-1} 1 - \frac{kp+kq+N_a}{M_{j+1}}\right)^{-1}$$

We prove the above for $\lambda = 1$, since the idea for the $p = q$ is the same.

Proof. For $\lambda = 1$, our recurrence is given by

$$X_{t+1} = X_t + \frac{1}{M_{t+1}}[\alpha N_a + kq_t - (kq_t + kp_t + N_a)X_t] + \frac{\mu_t}{M_{t+1}}$$
$$\implies \mathbb{E}[X_{t+1}|\mathscr{F}_t] = X_t\left(1 - \frac{kp+kq+N_a}{M_{t+1}}\right) + \frac{\alpha N_a + kq}{M_{t+1}}$$

Thus,

$$Z_T = X_T\left(\prod_{i=0}^{T-1} 1 - \frac{kp+kq+N_a}{M_{i+1}}\right)^{-1} - \sum_{i=0}^{T-1} \frac{kq+\alpha N_a}{M_{i+1}}\left(\prod_{j=0}^{i} 1 - \frac{kp+kq+N_a}{M_{j+1}}\right)^{-1}$$

is an $\{\mathscr{F}_t\}_t$-martingale.

One can see that $|Z_{t+1} - Z_t| \leq (3k+1+N_a)\prod_{i=1}^{t} \frac{M_i}{M_i - k(p+q)}$.

By the Azuma Hoeffding inequality (see 5.3.4), given $\varepsilon > 0$,

$$\mathbb{P}(|Z_T - Z_0| > \varepsilon) < \mu$$

where

$$\mu = 2\exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{T} c_i^2}\right)$$

and

$$\begin{aligned} c_t &= |Z_t - Z_{t-1}| \\ &= X_t\left(\prod_{i=0}^{t-1} 1 - \frac{kp+kq+N_a}{M_{i+1}}\right)^{-1} - X_{t-1}\left(\prod_{i=0}^{t-2} 1 - \frac{kp+kq+N_a}{M_{i+1}}\right)^{-1} \\ &- \frac{kq+\alpha N_a}{M_t}\left(\prod_{j=0}^{t-1} 1 - \frac{kp+kq+N_a}{M_{j+1}}\right)^{-1} \end{aligned}$$

For ease of notation, we write

$$Z_T \overset{\varepsilon}{\approx} Z_0 \text{ with probability } \mu$$

Thus, with probability μ, we have

$$X_T \overset{\varepsilon}{\approx} X_0 \prod_{i=0}^{T-1}\left(1 - \frac{kp+kq+N_a}{M_{i+1}}\right) + \sum_{i=0}^{T-1} \frac{kq+\alpha N_a}{M_{i+1}} \prod_{j=i+1}^{T-1}\left(1 - \frac{kp+kq+N_a}{M_{j+1}}\right)$$

For large T, we have

$$\begin{aligned} \prod_{i=0}^{T-1}\left(1 - \frac{kp+kq+N_a}{M_{i+1}}\right) &= \exp\left(\sum_{i=0}^{T-1} log(1 - \frac{kp+kq+N_a}{M_{i+1}})\right) \\ &\approx \exp\left(-\sum_{i=0}^{T-1} \frac{kp+kq+N_a}{M_{i+1}}\right) \approx \exp\left(-\int_0^{T-1} \frac{(kp+kq+N_a)di}{M_0 + (i+1)N_a}\right) \\ &= \left(\frac{T+1+M_0/N_a}{1+M_0/N_a}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)} \end{aligned}$$

By a similar method, we also get

$$\sum_{i=0}^{T-1} \frac{1}{M_{i+1}} \prod_{j=i+1}^{T-1} \left(1 - \frac{kp+kq+N_a}{M_{i+1}}\right)$$
$$\approx \int_0^{T-1} (T+1+M_0/N_a)^{-\left(\frac{kp+kq}{N_a}+1\right)} \frac{dj}{(j+2+M_0/N_a)^{-\left(\frac{kp+kq}{N_a}+1\right)}(M_0+(j+1)N_a)}$$
$$\approx \frac{(T+1+M_0/N_a)^{-\left(\frac{kp+kq}{N_a}+1\right)}}{N_a} \int_0^{T-1} \frac{dj}{(j+1+M_0/N_a)^{-\left(\frac{kp+kq}{N_a}\right)}}$$
$$= \frac{(T+1+M_0/N_a)^{-\left(\frac{kp+kq}{N_a}+1\right)}}{kp+kq+N_a}((T+M_0/N_a)^{\frac{kp+kq}{N_a}+1} - (1+M_0/N_a)^{\frac{kp+kq}{N_a}+1})$$

Using these two approximations, we obtain

$$X_T \overset{\varepsilon}{\approx} X_0 \left(\frac{T+1+M_0/N_a}{1+M_0/N_a}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)} + \left(\frac{kq+\alpha N_a}{kp+kq+N_a}\right)\left(1 - \left(\frac{T+1+M_0/N_a}{1+M_0/N_a}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)}\right)$$

with probability μ, where

$$\mu = 2\exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{T} c_i^2}\right)$$

where

$$c_t = |Z_{t+1} - Z_t| \leq (3k+1+N_a)\prod_{i=1}^{T} \frac{M_i}{M_i - k(p+q)}$$

□

4.3.2 Optimal Strategies

In this section we compare the two extreme influencing strategies $\mathscr{S}_F$ and $\mathscr{S}_L$. For $\mathscr{S}_F$, we influence the first bT slots (where $b \in [0,1]$) with probabilities $p_t = \tilde{p}$ and $q_t = \tilde{q}$. Since the solution of the recursion tracks the solution of the ODE, we solve the ODE and compare the final fraction of people with opinion "Yes" by integrating the appropriate ODEs for the given slots of time. We assume throughout that the influencing agency is rational and therefore $\tilde{q} > \tilde{p}$.

Let X_T^L and X_T^F be the ODE solutions for last and first bT respectively. Define the following:

- $x = k((\tilde{p}+\tilde{q})-(p+q))$
- $A(x) = \frac{k\tilde{q}+\alpha N_a}{k(\tilde{p}+\tilde{q})+N_a} - \frac{kq+\alpha N_a}{kp+kq+N_a}$
- $\chi = \frac{T}{1+M_0/N_a}$

Let $D(x) = X_T^L - X_T^F$. Then,

$$\begin{aligned} D(x) &= X_T^L - X_T^F \\ &= A\left[1 - \left(\frac{\chi+1}{(1-b)\chi+1}\right)^{-\left(\frac{x+kp+kq}{N_a}+1\right)} - \left(\frac{\chi+1}{b\chi+1}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)}\right] \\ &+ \left(x_0 - \frac{kq+\alpha N_a}{kp+kq+N_a}\right)((1-b)\chi+1)^{x/N_a}(\chi+1)^{-\left(\frac{x+kp+kq}{N_a}+1\right)} \\ &- \left(x_0 - \frac{k\tilde{q}+\alpha N_a}{k\tilde{p}+k\tilde{q}+N_a}(b\chi+1)^{-x/N_a}(\chi+1)^{-\left(\frac{kp+kq}{N_a}+1\right)}\right) \end{aligned}$$

Theorem 4.3.2. *Given $x = 0$ and $\tilde{q} > q$, $\tilde{p} < p$, for $\alpha = 1/2$, we have $\mathscr{S}_L >> \mathscr{S}_F$.*

Proof. Using stochastic approximation, we can analyse $D(0)$. One can see that

$$\begin{aligned} D(0) &= A\left[1 - \left(\frac{\chi+1}{(1-b)\chi+1}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)} - \left(\frac{\chi+1}{b\chi+1}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)} + (\chi+1)^{-\left(\frac{kp+kq}{N_a}+1\right)}\right] \\ &= A \times F_b(\chi) \end{aligned}$$

where $F_b(\chi)$ attains a minima at $b = 0$ or $b = 1$. In particular, $F_0(\chi) = F_1(\chi) = 0$. □

To analyse the optimality in general, we have the following cases:

(a) For $M_0 << T$ or $M_0 >> T$ and $M_0 << N_a$ or $M_0 \approx N_a$ then $\chi \approx \gamma T$ for some constant γ (for example $\gamma = 1/2$ for $M_0 \approx N_a$). γ is very small for large T and we can take $\chi \approx T$.

(b) For $M_0/N_a >> T$, $\chi \approx \varepsilon$ for ε very small.

(c) For $M_0 \approx T >> N_a$, $\chi \approx 1$.

Theorem 4.3.3. *For $k = 1$ and for χ very large, i.e., for $T >> 1 + M_0/N_a$, both strategies are optimal at $A(x) = 0$ (denote as x_A). For $x \in [-2, x_A)$, $\mathscr{S}_L >> \mathscr{S}_F$ and for $x \in (x_A, 2]$, $\mathscr{S}_L << \mathscr{S}_F$.*

(sketch). Take $\chi = L$ where L is very large, such that $L+1 \approx L$ and $cL+1 \approx cL$ for any real c. Then,

$$D(x) \approx A\left[1 - \left(\frac{1}{1-b}\right)^{-\left(\frac{x+kp+kq}{N_a}+1\right)} - \left(\frac{1}{b}\right)^{-\left(\frac{kp+kq}{N_a}+1\right)}\right]$$

Given that x must lie within $[-2,2]$ for $k=1$, one can see that $D(x)=0$ when $A(x)=0$, and that $D(x)$ is decreasing with an increasing x. □

Theorem 4.3.4. *For χ very small, i.e., for $T << 1 + M_0/N_a$, both strategies are optimal.*

Proof. Take $\chi = \varepsilon$, where ε is very small. Then, $1+\varepsilon \approx 1$, and $b\varepsilon \approx \varepsilon$. Then,

$$D(x) \approx -A + A = 0$$

□

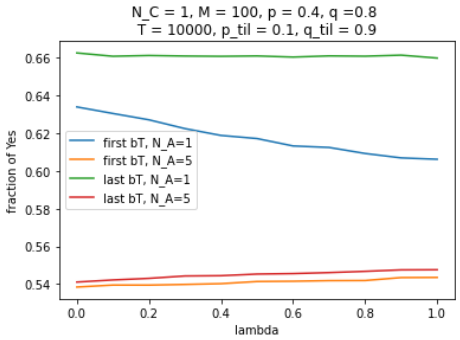
4.4 Simulations

For all simulations, we fix the following.

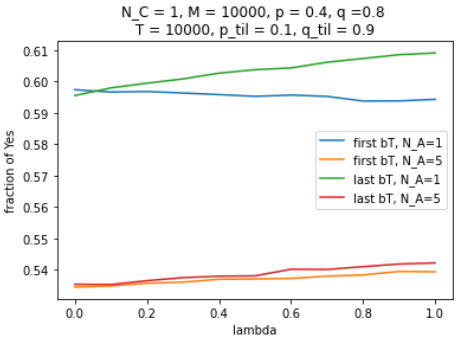
- $T = 10,000$
- $\tilde{p} = 0.1$ and $\tilde{q} = 0.9$
- $\alpha = 0.5$
- $b = 0.4$

Through simulations, we study the effect of the parameters N_a and N_c, as well as observe how close the simulated opinion evolution is to the ODE solution obtained by stochastic approximation. We then make some observations and conjectures, and compare them to the main result of [SKG+20].

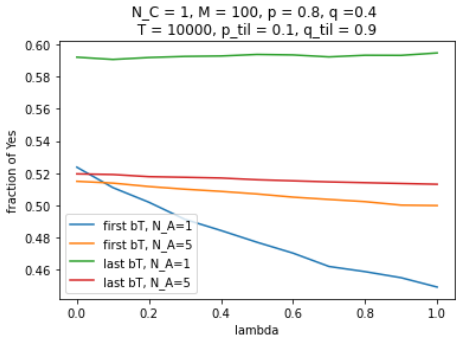
4.4.1 Effect of N_a



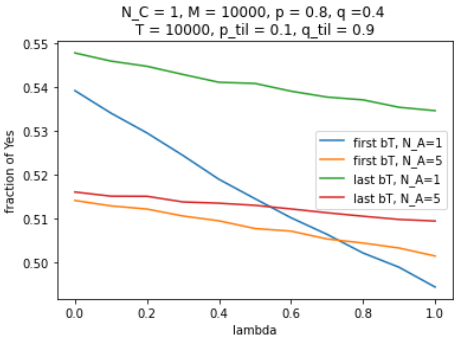
(a) $p < q$, $M = 100$



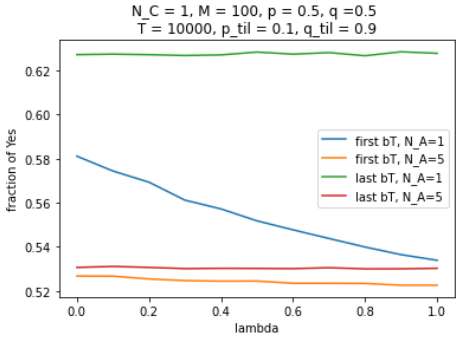
(b) $p < q$, $M = 10000$



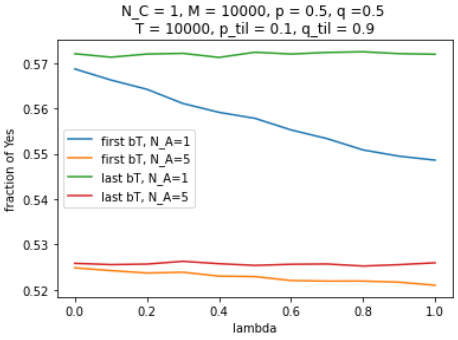
(c) $p > q$, $M = 100$



(d) $p > q$, $M = 10000$



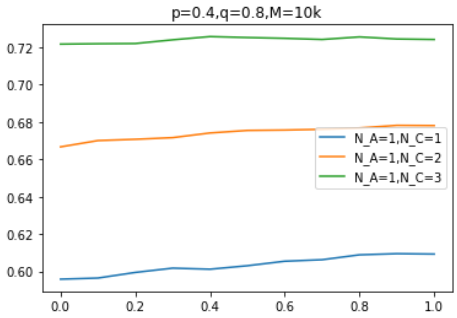
(e) $p = q$, $M = 100$



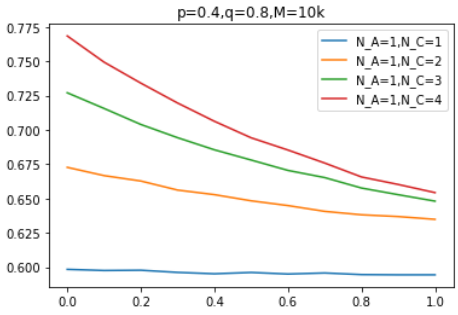
(f) $p = q$, $M = 10000$

Figure 4.1: Studying the effect of N_a on influencing strategies across varying λ

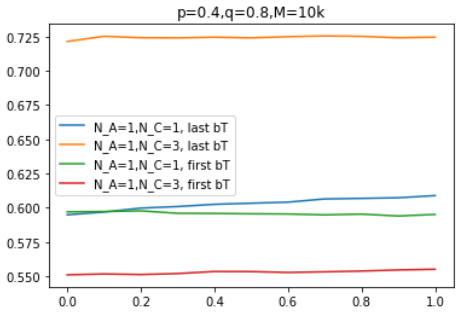
4.4.2 Effect of N_c



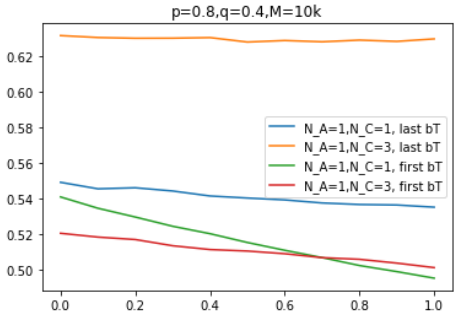
(a) $p < q$, $M = 10000$, last bT



(b) $p < q$, $M = 10000$, first bT



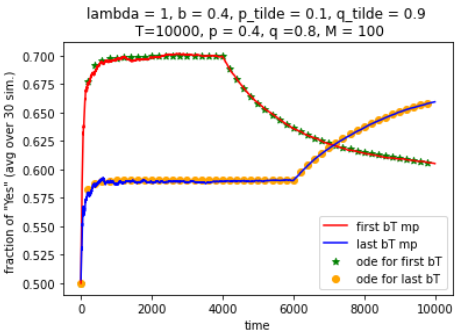
(c) $p < q$, $M = 10k$



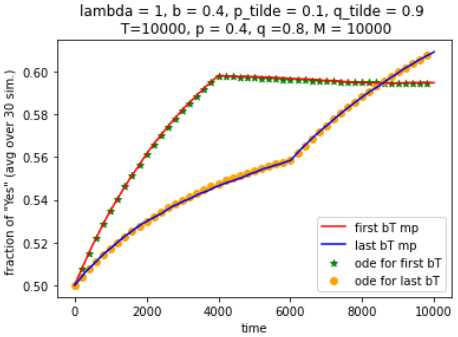
(d) $p > q, M = 10k$

Figure 4.2: Studying the effect of N_c on influencing strategies across varying λ

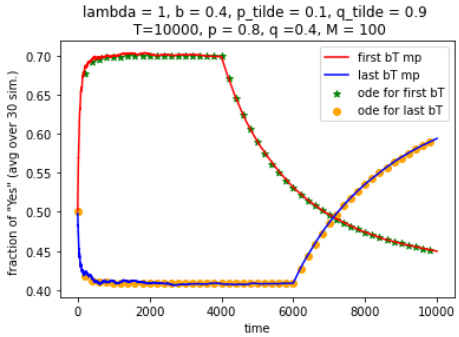
4.4.3 Stochastic Approximation and ODE



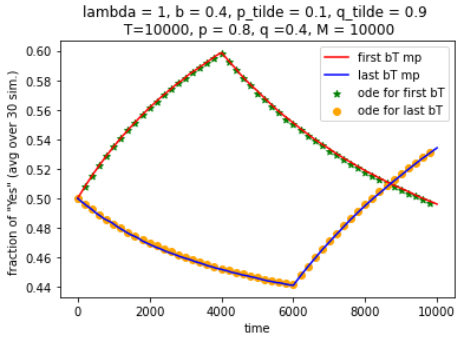
(a) $p < q$, $M = 100$



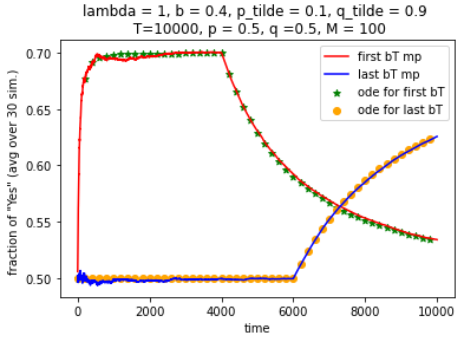
(b) $p < q$, $M = 10000$



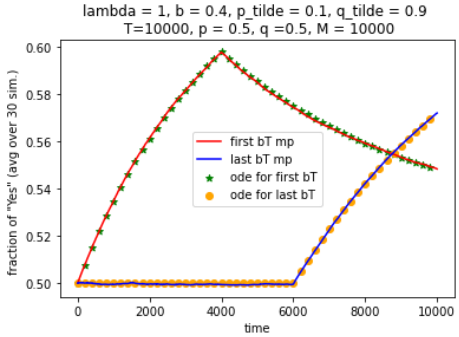
(c) $p > q$, $M = 100$



(d) $p > q$, $M = 10000$



(e) $p = q$, $M = 100$



(f) $p = q$, $M = 10000$

Figure 4.3: A comparison of the solution of the ODE and the simulated population

4.4.4 Inferences

We analyse the simulated plots and compare our conjectures to Theorem 1 of [SKG+20].

p, q	M value	Results for fixed population from Theorem 1 of [SKG+20]	Observations for growing population with $N_a = N_c = 1$
0.4, 0.8	100	Both optimal for $\lambda = 0$, $\mathcal{S}_F$ for $\lambda > 0$	$\forall\ \lambda$, $\mathcal{S}_L$ is optimal
0.4, 0.8	10,000	$\exists\ \lambda^*$ s.t. $\mathcal{S}_F$ optimal for $\lambda < \lambda^*$, $\mathcal{S}_L$ for $\lambda > \lambda^*$, both at λ^*	similar result with $\lambda^* \approx 0.1$
0.5, 0.5	100	$\mathcal{S}_L$ optimal for $\lambda > 0$, both for $\lambda = 0$	$\mathcal{S}_L$ is optimal
0.5, 0.5	10,000	$\mathcal{S}_L$ optimal for $\lambda > 0$, both for $\lambda = 0$	$\mathcal{S}_L$ is optimal, very close to $\mathcal{S}_F$ at $\lambda = 0$
0.8, 0.4	100	$\mathcal{S}_L$ is optimal	$\mathcal{S}_L$ is optimal
0.8, 0.4	10,000	$\mathcal{S}_L$ is optimal	$\mathcal{S}_L$ is optimal

We additionally make the following conjectures

- As N_a increases, the difference between the two strategies ($\mathscr{S}_F$ and $\mathscr{S}_L$) starts reducing. This holds true for all values of p, q and M, with $N_c = 1$. Moreover, N_a does not change optimality.

  Reasoning: *Since N_a individuals are added with opinion "1" or "0" with probability α and $1-\alpha$ resp., their effect on fraction of individuals of a given opinion drastically overpowers the effect of changing the opinion of one individual as N_a increases. Thus, for N_a sufficiently larger than N_c, the fraction of "1" would converge closer to α.*

- The smaller the initial size M of the population, the larger is the difference between $\mathscr{S}_F$ and $\mathscr{S}_L$.

  Reasoning: *A smaller initial population is likely to converge faster than a larger initial population, since the effect of adding an individual and changing an opinion is more significant on a smaller M.*

4.5 Future Scope

Existing literature studies Opinion Dynamics of a fixed size population with an underlying graph structure. For a growing population, to introduce an underlying graph, the graph itself must be dynamic, which is a property of the Preferential attachment graph. The opinion of every individual can be thought of as their fitness.

In our work, at every step an individual is chosen uniformly at random. There have been studies on models where the individual (who changes their opinion at time n) is chosen according to an underlying Markov chain on the graph. The asymptotics of the opinion evolution then depend on the stationary distribution of the underlying Markov chain, or the Pagerank of that chain.

Chapter 5

Appendix

5.1 Pólya Urn Process

A Pólya-Eggenberger urn, or simply the Pólya urn model, is an urn process consisting of an urn with balls of two colours. At every time step, a ball from the urn is picked uniformly at random, and put back into the urn along with a reinforcement of the same colour. Thus, the process appropriately gets the property *"rich get richer"*. A key result in the theory of Pólya urns is the distribution of number of successes, that is, the number of times a ball of a particular colour is picked in n draws. The following theorem comes from Theorem 3.1, chapter 3, of the text [Mah].

Theorem 5.1.1 (Distribution of number of successes)**.** *For a Pólya urn process with reinforcement matrix given by*

$$\mathscr{R} = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$$

with an initial number of b_0 black balls and w_0 white balls, the probability $\mathscr{P}_{w_0,b_0,s}(n,k)$ of drawing k black balls in n time steps is given by

$$\mathscr{P}_{w_0,b_0,s}(n,k) = \binom{n}{k} \frac{b_0(b_0+s)\ldots(b_0+(k-1)s)w_0(w_0+s)\ldots(w_0+(n-k-1)s)}{(b_0+w_0)(b_0+w_0+s)\ldots(b_0+w_0+(n-1)s)}$$

Theorem 3.2 of the same text shows that the fraction of successes converges in distribution to the **beta distribution**.

5.2 Stochastic Approximation

Theorem 5.2.1. *Consider the following recurrence relation*

$$x(n+1) = x(n) + a(n)\left[h(x(n)) + M(n+1)\right],\ n \geq 0$$

such that

1. $\{a(n)\}$ *is a positive step-size sequence satisfying*

$$\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty.$$

2. $h : \mathbb{R}^d \to \mathbb{R}^d$ *is Lipschitz.*

3. $\{M_n\}_{n\geq 0}$ *is a square-integrable Martingale difference sequence with respect to a suitable filtration.*

4. $\sup_n \|x(n)\| < \infty$.

Consider the ODE

$$\dot{x} = h(x(t))$$

Then, asymptotically with probability 1, the iterates 5.2.1 converge almost surely to the stable equilibria of the solutions of the above ODE.

The result comes from Borkar (2008).

5.3 Concentration Inequalities

Theorem 5.3.1 (Hoeffding Inequality)**.** *Let* $\{X_i\}_{i=1}^t$ *be independent random variables such that* $\mathbb{P}(X_k \in [a_k, b_k]) = 1$ *for some finite real* a_k *and* b_k*. Let* $X = \sum_{i=1}^{t} X_i$*. Then,*

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq \delta\right) \leq 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^{t}(a_k - b_k)^2}\right\}$$

Theorem 5.3.2 (Chernoff Bound)**.** *Let* $\{X_i\}_{i=1}^n$ *be random variables and* $X = \sum_{i=1}^{n} X_i$*. Then, for all* $t > 0$*, we have*

$$\mathbb{P}(X \geq a) \leq e^{-ta}\mathbb{E}\left[\prod_i e^{tX_i}\right]$$

For X_i independent, we get

$$\mathbb{P}(X \geq a) \leq \min_{t>0} e^{-ta} \prod_i \mathbb{E}[e^{tX_i}]$$
$$\mathbb{P}(X \leq a) \leq \min_{t>0} e^{ta} \prod_i \mathbb{E}[e^{-tX_i}]$$

The above bound is obtained by applying the Markov inequality to e^{tX}.

Definition 5.3.3. *Let $X = \{X_t\}_{t\in I}$ be a real-valued and $\{\mathscr{F}_t\}_{t\geq 0}$-adapted stochastic process with $\mathbb{E}[|X_t|] < \infty$ for all $t \in I$, where $\{\mathscr{F}_t\}_{t\geq 0}$ is a filtration. Then, X is said to be a martingale with respect to $\{\mathscr{F}_t\}_{t\geq 0}$ if*

$$\mathbb{E}[X_t|\mathscr{F}_s] = X_s$$

for all $t > s$.

Theorem 5.3.4 (Azuma-Hoeffding inequality)**.** *Suppose $\{X_k\}_{k\geq 0}$ is a martingale such that almost surely*

$$|X_k - X_{k-1}| \leq c_k$$

Then, for all $N \in \mathbb{N}$ and $\varepsilon > 0$, we have

$$\mathbb{P}(|X_N - X_0| \geq \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2\sum_{k=1}^{N} c_k^2}\right)$$

Lemma 5.3.5. *Let $\{a_t\}_{t\geq 0}, \{b_t\}_{t\geq 0}$ and $\{c_t\}_{t\geq 0}$ be three real sequences such that*

$$a_{t+1} = a_t\left(1 - \frac{b_t}{t}\right) + c_t$$

with $\lim_{t\to\infty} b_t = b \geq 0$ and $\lim_{t\to\infty} c_t = c$. Then,

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}$$

Proof.

$$\begin{aligned}\frac{a_{t+1}}{t+1} - \frac{c}{1+b} &= \frac{(1-b_t/t)a_t + c_t}{1+t} - \frac{c}{1+b} \\ &= \left(\frac{a_t}{t} - \frac{c}{1+b}\right)\left(1 - \frac{1+b_t}{1+t}\right) + \frac{(1+b)c_t - (1+b_t)c}{(1+b)(t+1)}\end{aligned}$$

Take $s_t = \left|\frac{a_t}{t} - \frac{c}{1+b}\right|$. Then,

$$s_{t+1} \leq s_t\left|1 - \frac{1+b_t}{1+t}\right| + \left|\frac{(1+b)c_t - (1+b_t)c}{(1+b)(t+1)}\right|$$

Since $\lim_{t\to\infty} b_t = b$ and $\lim_{t\to\infty} c_t = c$, we can say that $\forall \varepsilon > 0$, $\exists\ t_0$ such that $\forall t > t_0$, $|(1+b)c_t - (1+b_t)c| < \varepsilon$. We also have that $b_t > b/2\ \forall t > T$ for some T, and thus,

$$
\begin{aligned}
s_{t+1} &\leq s_t \left|1 - \frac{1+b_t}{1+t}\right| + \left|\frac{(1+b)c_t - (1+b_t)c}{(1+b)(t+1)}\right| \\
\implies s_{t+1} &< s_t \left(1 - \frac{1+b/2}{1+t}\right) + \frac{\varepsilon}{(1+b)(t+1)} \\
\implies s_{t+1} - \varepsilon &< s_t \left(1 - \frac{1+b/2}{1+t}\right) - \varepsilon \left(1 - \frac{1+b/2}{(1+b)(1+b/2)(t+1)}\right) \\
&< s_t \left(1 - \frac{1+b/2}{1+t}\right) - \varepsilon \left(1 - \frac{1+b/2}{t+1}\right) \\
&< (s_t - \varepsilon) \left(1 - \frac{1+b/2}{t+1}\right) \\
&< (s_1 - \varepsilon) \prod_{i=1}^{t} \left(1 - \frac{1+b/2}{i+1}\right)
\end{aligned}
$$

For $b > 0$, the product goes to 0 as $t \to \infty$. For $b = 0$, we have

$$
\prod_{i=1}^{t} \left(1 - \frac{1}{i+1}\right) = \prod_{i=1}^{t} \left(\frac{i}{i+1}\right) = \frac{1}{t!} \xrightarrow{t\to\infty} 0
$$

□

The above result and proof are from Chapter 3.3 of . We have a minor tweak of $b \geq 0$ instead of $b > 0$ in this result.

Bibliography

[Agu09] Rafik Aguech, *Limit Theorems for Random Triangular URN Schemes*, Journal of Applied Probability **46** (2009), no. 3, 827–843 (en).

[AL06] Konstantin Avrachenkov and Dmitri Lebedev, *PageRank of Scale-Free Growing Networks*, Internet Mathematics **3** (2006), no. 2, 207–231 (en).

[Ath07] K. B. Athreya, *Preferential Attachment Random Graphs with General Weight Function*, Internet Mathematics **4** (2007), no. 4, 401–418 (en).

[BB01] G Bianconi and A.-L Barabási, *Competition and multiscaling in evolving networks*, Europhysics Letters (EPL) **54** (2001), no. 4, 436–442 (en).

[BCDR07] Christian Borgs, Jennifer Chayes, Constantinos Daskalakis, and Sebastien Roch, *First to Market is not Everything: an Analysis of Preferential Attachment with Fitness*, arXiv:0710.4982 [cs, math] (2007) (en), arXiv: 0710.4982.

[DM08] Steffen Dereich and Peter Morters, *Random networks with sublinear preferential attachment: Degree evolutions*, arXiv:0807.4904 [math] (2008) (en), arXiv: 0807.4904.

[ER02] G. Ergun and G. J. Rodgers, *Growing Random Networks with Fitness*, Physica A: Statistical Mechanics and its Applications **303** (2002), no. 1-2, 261–272 (en), arXiv: cond-mat/0103423.

[FM] Y. Feng and H. Mahmoud, *Dynamic Pólya-Eggenberger urns*, Statistics and Probability letters (en).

[GMS20] Anmol Gupta, Sharayu Moharir, and Neeraja Sahasrabudhe, *Influencing Opinion Dynamics in Networks with Limited Interaction*, arXiv:2002.00664 [cs, eess] (2020) (en), arXiv: 2002.00664.

[GvdHW17] Alessandro Garavaglia, Remco van der Hofstad, and Gerhard Woeginger, *The Dynamics of Power laws: Fitness and Aging in Preferential Attachment Trees*, Journal of Statistical Physics **168** (2017), no. 6, 1137–1179 (en).

[KR01] P. L. Krapivsky and S. Redner, *Organization of growing random networks*, Physical Review E **63** (2001), no. 6, 066123 (en).

[KRL00] P L Krapivsky, S Redner, and F Leyvraz, *Connectivity of Growing Random Networks*, PHYSICAL REVIEW LETTERS **85** (2000), no. 21, 4 (en).

[Mah] Hosam M. Mahmoud, *Pólya urn models*, CRC Press, Taylor and Francis Group.

[OS05] Roberto Oliveira and Joel Spencer, *Connectivity transitions in networks with super-linear preferential attachment*, arXiv:math/0510446 (2005) (en), arXiv: math/0510446.

[SKG+20] Arunabh Saxena, Bhumesh Kumar, Anmol Gupta, Neeraja Sahasrabudhe, and Sharayu Moharir, *Influencing Opinions of Heterogeneous Populations over Finite Time Horizons*, arXiv:1905.04448 [cs] (2020) (en), arXiv: 1905.04448.

[TGP20] Thibaud Trolliet, Frédéric Giroire, and Stéphane Pérennes, *A Random Growth Model with any Real or Theoretical Degree Distribution*, arXiv:2008.03831 [cs] (2020) (en), arXiv: 2008.03831.

[vdH14] Remco van der Hofstad, *Random graphs and complex networks, volume i*, 2014.

[Ver] Roman Vershynin, *High-Dimensional Probability* (en).

[Ye] Wenxing Ye, *On PageRank Algorithm and Markov Chain Reduction*, 4 (en).