Chapter 4

Models for the Web Graph

There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world.

–Nikolai Lobachevsky

All models are wrong, but some models are useful.

–G.E.P. Box

4.1. Introduction

While graph theory is usually studied in the abstract and is a fascinating subject in its own right, one of its long-recognized applications is to model networks in the real world. The web graph W is one such real-world network, and is our focus. In modelling the web, the many parameters and complex phenomena at work in its evolution are distilled into a simplified picture. Good models for W usually have only a few parameters, and must capture salient features of W. The balance between simplicity of design and the difficulties of analyzing the model makes the subject of modelling W challenging.

In this chapter we will discuss some of the key models for the web graph W. Our approach will be to consider *stochastic models*: that is, models where graphs are generated over an infinite sequence of discrete time-steps via certain probabilistic rules. Early models of W were often posed and analyzed in a non-rigorous fashion. The present challenge is to design mathematically

rigorous models that simulate one or more of the features observed in W (as described in Chapter 2), and can be rigorously analyzed. As we will see, this challenge has been met by an increasing number of rigorous models. To analyze these models, the techniques from Chapter 3 come into play, as well as new ones that we describe as they are needed.

What features make a good web graph model? There is no definitive answer to this question, only a consensus based on observed properties of W. The following is a partial list of desirable properties that graphs generated by a web graph model should possess, based on the observed properties of W described in Chapter 2.

(1) *On-line property.* Graphs generated by the model change with time, so both the number of vertices and the number of edges change with time.

(2) *Power law degree distribution.* A.a.s. the degree distribution of graphs generated by the model follows a power law with exponent $\beta > 2$.

(3) *Small world property.* The model a.a.s. generates sparse graphs with "low" diameter and average distance. For example, the diameter should be a.a.s. approximately $\log t$ if there are t vertices, while the average distance should be approximately $\log \log t$.

Additional, but less frequently studied, desirable properties for a model of W include higher values of the clustering coefficient when compared to a random graph of similar size and order, a larger number of bipartite cliques when compared to a random graph of similar size and order, and sparse cuts. For simplicity, we will focus on properties (1), (2), and (3) above, but we will mention some of these other properties along the way as we describe various models.

Pioneering work on random graphs was first done by Erdős and Rényi [**98, 99**], as discussed in Chapter 3. We begin by recalling the random graph $G(n, p)$. We are given n vertices and a fixed real number parameter $p \in (0, 1)$. For each of the $\binom{n}{2}$ many distinct pairs of vertices, add an edge between them independently with probability p. The probability space $G(n, p)$, in a certain sense, is static or *off-line*: the number of vertices is fixed. Although usually n is taken as very large, and the number of edges is viewed as being variable over time, the number of vertices in $G(n, p)$ is off-line. Few techniques or models were available before the late 1990's for on-line random graph models. Further, from Theorem 3.11 the degrees of vertices are binomially distributed. Hence, based on items (1) and (2) above, $G(n, p)$ is not appropriate as a model of the web graph W (after all, the study of random graphs predates the inception of the internet by several

decades). Nevertheless, random graphs supply the mathematical subtext for these new models, and many of the techniques used to analyze them are also useful for models of W.

Over the last decade, a large number of rigorous models for the web graph W have been proposed. Such models deepen our understanding of the generative mechanisms driving the evolution of W, and provide insight into superficially unrelated properties observed in the web.

We make two caveats. First, we focus on a handful of the most influential models, and on the properties (1), (2), and (3) described above. As our goal is an introduction to models of W, we do not claim to survey all the models in the literature. Models of W (and many other self-organizing networks) have been analyzed with regards to several other properties, such as their eigenvalues [**78, 79, 161**], vulnerability to attack [**36**], orders of their connected components [**75**], and spread of viruses and worms on W [**25**]. See Chapter 7 for a discussion of some of these topics. The second caveat is that, since this is a mathematics text, we focus only on rigorous models and analysis.

In Section 4.2.1 we consider preferential attachment models for W, and give a rigorous analysis of the degree distribution of one of these models. Other models are described, including the copying model (Section 4.2.5), growth-deletion models (Section 4.2.6), geometric models (Section 4.2.7), and an off-line model (Section 4.2.8). We finish with a description of some challenges faced in future modelling of W.

4.2. On-Line Web Graph Models

As discussed in the previous section, our emphasis is on *on-line* web graph models: that is, models whose vertex set increases in cardinality over time. This approach is the most desirable one for models of W, which is, after all, a dynamic graph. A central idea in all models is to consider both approximate results and asymptotic behaviour. The rationale for both is that W is a massive graph with a large number of vertices and edges, and on average, small changes make little difference in the structure of the overall graph. Random asymptotic techniques are therefore most suitable for analyzing these models.

To simplify notation, we supply the following framework for all the on-line models we present. The model will always possess a finite set of real number parameters (the fewer the better), and has a fixed finite graph H as an additional parameter. The model generates by some random graph process a sequence of finite graphs G_t indexed by $(t : t \in \mathbb{N})$. Unless otherwise stated, for all $t \in \mathbb{N}$, we have that

(1) $G_0 \cong H$;
(2) G_t is an induced subgraph of G_{t+1};
(3) $|V(G_{t+1})| = |V(G_t)| + 1$.

In all the on-line models we consider, the graphs G_t are defined inductively. In the inductive step, the unique vertex in $V(G_{t+1})\backslash V(G_t)$ is referred to as the *new vertex*, written v_{t+1}, and the vertices of $V(G_t)$ are the *existing vertices*. We note that the choice of H usually has no effect on the value of the power law exponent β, while the choice of real number parameters does generally affect β. Further, the number of edges in $E(G_{t+1})\backslash E(G_t)$ is usually a constant parameter. However, this is not always the case: the numbers of edges and vertices in G_t may be random variables.

For $k, t \geq 0$ integers, define $N_{k,t}$ to be the number of vertices of degree k at time t. Then $N_{k,t}$ itself is a random variable. Since the number of vertices in G_t is usually approximately t, most authors study the proportion $\frac{N_{k,t}}{t}$. The most common technique for proving that $\frac{N_{k,t}}{t}$ follows a power law involves first computing $\mathbb{E}\left(\frac{N_{k,t}}{t}\right)$, and then proving that $\frac{N_{k,t}}{t}$ does not deviate too far from $\mathbb{E}\left(\frac{N_{k,t}}{t}\right)$. Authors prove that the random variable *concentrates* on its expected value. As we will see, both the computation of the expected value of $\frac{N_{k,t}}{t}$ and deriving concentration around the expected value are non-trivial problems for models of W.

4.2.1. Preferential attachment models. Arguably the most important web graph models are ones incorporating some form of preferential attachment. The first evolving graph model explicitly designed to model W was given by Barabási and Albert [**21**]. The idea behind their model is an intuitively pleasing one: new vertices are more likely to join to existing vertices with high degree. In a slogan, *the rich get richer*. This model is now referred to as an example of a *preferential attachment* (or *PA*) *model.* As we will discuss in detail, PA models generate graphs with properties quite different from those of $G(n, p)$. See Figures 4.1 and 4.2 to compare and contrast graphs generated by these two models.

Barabási and Albert gave a heuristic description and analysis of their PA model (using mean field theory from physics), and concluded that it generates graphs whose in-degree distribution follows a power law with exponent $\beta = 3$. Although their proof was not rigorous, their important work set the stage for most of the mathematics regarding the modelling of W to come.

4.2.2. The LCD PA model. The first rigorous analysis of a PA model was given in Bollobás, Riordan, Spencer, and Tusnády [**34**]. Their model is called the *Linearized Chord Diagram* or *LCD* model, since an equivalent

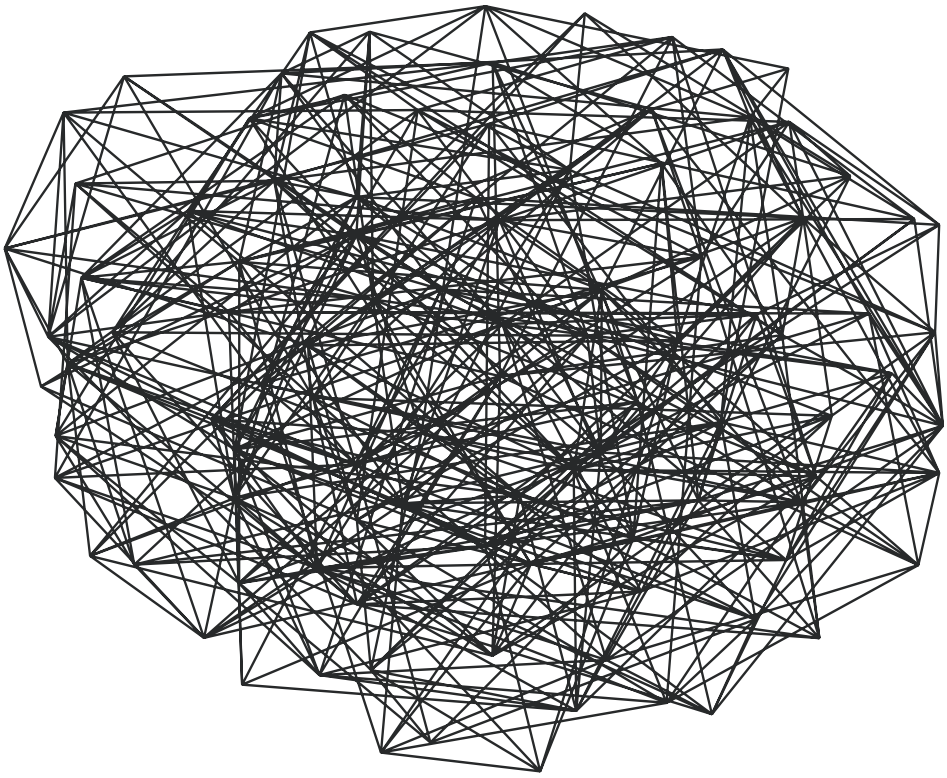

Figure 4.1. A graph with 100 vertices and edges drawn with probability $\frac{1}{2}$.

formulation of the model is via random pairings on a fixed finite set of integers, as we will describe below. The sole parameter of this model is a positive integer m, where H is a copy of K_1 with a single loop. We first describe the model in the case $m = 1$. To form G_t from G_{t-1} add a single edge from v_t to v_i, where the vertex v_i is chosen at random from the existing vertices, with

$$\mathbb{P}(i = s) = \begin{cases} \frac{\deg_{G_{t-1}}(v_s)}{2t-1} & \text{if } 1 \leq s \leq t-1, \\ \frac{1}{2t-1} & \text{if } s = t. \end{cases} \tag{4.1}$$

Note that (4.1) gives a higher probability for an existing vertex to acquire a new edge if it has high degree. The graph G_t contains no non-trivial cycles, although self-loops are allowed. A similar version of this model was previously studied (in a different context) as *random recursive trees*; see [**34**] for further discussion.

If $m > 1$, then define the process $(G_t^m : t \geq 0)$ by first generating a sequence $(G_t : t \in \mathbb{N})$ of graphs using the case $m = 1$ on a sequence of

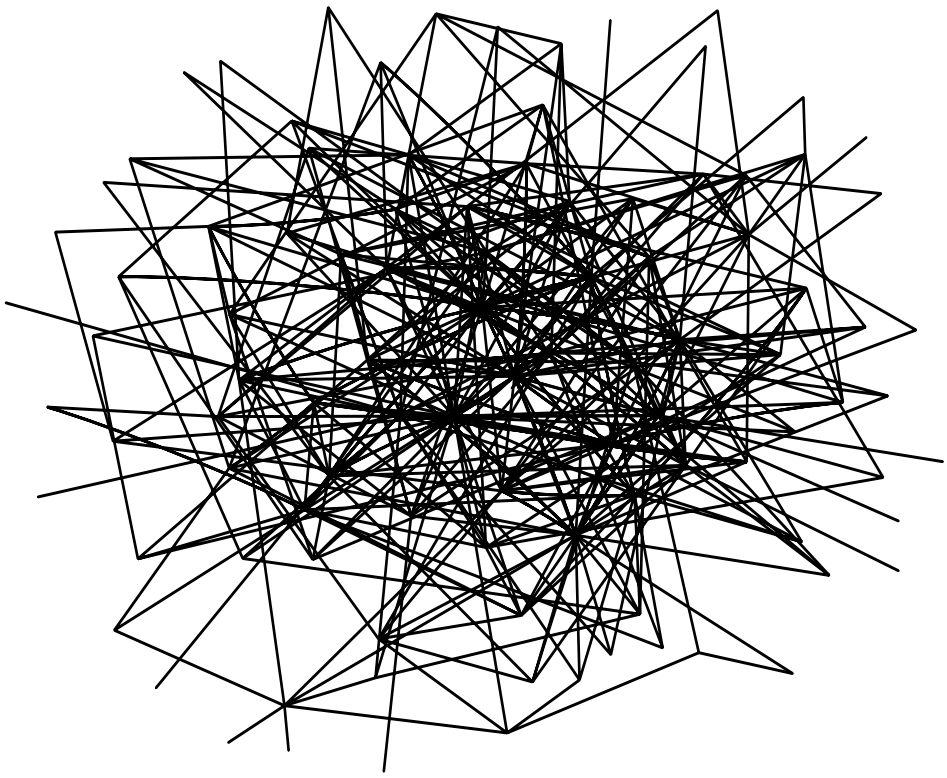

Figure 4.2. A graph generated by the preferential attachment model with 100 vertices and average degree close to the average degree of the graph in Figure 4.1.

vertices $(v'_i : i \in \mathbb{N} \setminus \{0\})$. The graph G_t^m is formed from G_{mt} by identifying the vertices $v'_{(i-1)m+1}, \ldots, v'_{im}$ to form v_i.

To analyze the model, the authors use the following geometric notion. Define a *linearized chord diagram* or *LCD* of order t to be a partition of $[2t]$ into n sets of distinct pairs. Then there are $\frac{(2t)!}{t!2^t}$-many LCD's on $[2t]$ (see Exercise 3).

We may identify an LCD with t-many semi-circular chords between $2n$ distinct points on the x-axis paired off in the upper half-plane of $\mathbb{R}^2$. Each chord has a left and right endpoint. For a fixed LCD L we may form an undirected graph $G(L)$ of order t as follows. To form vertex v_1, identify all endpoints up to and including the first right endpoint of all the chords. Proceed inductively to define the remaining vertices, so that the $(k+1)$th vertex is formed by identifying all endpoints up to and including the first right endpoint after vertex v_k. The chords may be viewed as (multiple) edges among the vertices. For an example, see Figure 4.3.

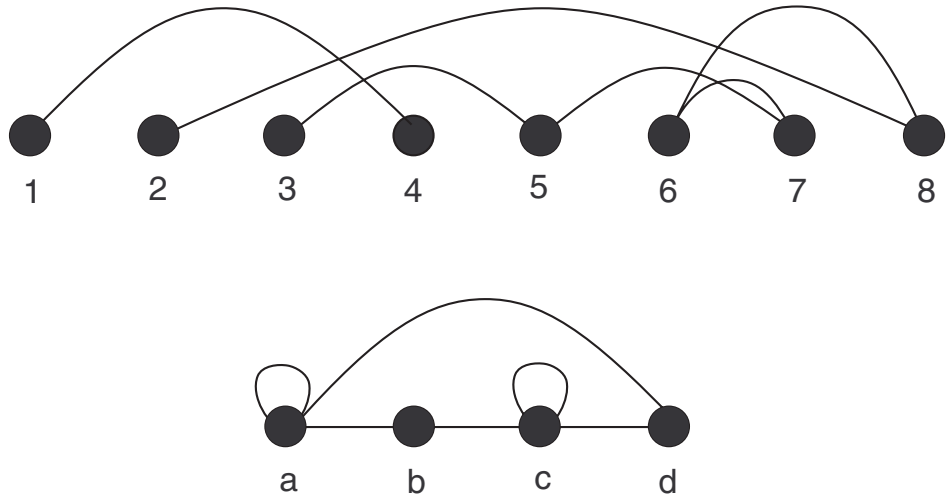

Figure 4.3. An LCD with the corresponding graph. The vertex a is identified with $1, 2, 3, 4$, the vertex b with 5, the vertex c with $6, 7$, and d with 8.

The connection between LCD's and the LCD model comes from the following theorem (whose proof is an exercise).

THEOREM 4.1. *Let $m = 1$. Suppose that an LCD L is chosen u.a.r. from the $\frac{(2t)!}{t!2^t}$ many LCD's on $[2t]$. Then the probability that vertices v_i and v_s are joined is given by (4.1).*

Bollobás et al. prove the following theorem.

THEOREM 4.2 ([**34**]). *In the LCD model, fix m a positive integer, and fix $\varepsilon > 0$. For k a non-negative integer, define*

$$\alpha_{m,k} = \frac{2m(m+1)}{(k+m)(k+m+1)(k+m+2)}.$$

Then a.a.s. for all k satisfying $0 \leq k \leq t^{1/15}$,

$$(1-\varepsilon)\alpha_{m,k} \leq \frac{N_{k,t}}{t} \leq (1+\varepsilon)\alpha_{m,k}. \tag{4.2}$$

Theorem 4.2 demonstrates that for large t, with high probability the degree distribution of graphs in the LCD model follows a power law with exponent $\beta = 3$ (formally justifying the conclusions derived in [**21**]). The reader will note that Theorem 4.2 is stated as a concentration result for degrees in the range $0 \leq k \leq t^{1/15}$; as remarked in [**34**], this may be extended to degrees $k > t^{1/15}$. An important observation is that the power law exponent $\beta = 3$ is independent of the choice of m. While $\beta = 3$ is in the correct range to model W, this restrictiveness of the power law exponent

can be viewed as a drawback of the LCD model. In Chapter 6, we explore how infinite limit graphs play a role in distinguishing different values of m.

The non-trivial part of the proof of Theorem 4.2 involves estimating $\mathbb{E}(N_{k,t})$ to be $\alpha_{m,k}t$. Rather than give a full proof of Theorem 4.2, we will prove a similar result for a closely related, but more easily stated PA model in Section 4.2.3.

THEOREM 4.3. *For all k satisfying $0 \leq k \leq t^{1/15}$, the sequence $(N_{k,t} : t \in \mathbb{N})$ converges in probability to $\mathbb{E}(N_{k,t})$.*

Using Theorem 4.3 and assuming that $\mathbb{E}(N_{k,t}) \sim \alpha_{m,k}t$ in the range $0 \leq k \leq t^{1/15}$, (4.2) follows. We give a proof of the concentration result in the case $m = 1$ (for the general case, see [**34**]), using the method of bounded differences as described in Chapter 3.

Proof of Theorem 4.3 *if* $m = 1$. Using notation from Chapter 3, consider the Doob martingale $(X_i : 0 \leq i \leq t)$ defined by $X_0 = \mathbb{E}(A)$ and let $X_i = \mathbb{E}[A|Z_1, \ldots, Z_i]$, where $A = N_{k,t}$ and $Z_i = G_i^m$ (we consider the graphs from earlier time-steps as random variables). Since a new vertex can affect the degrees of at most two existing vertices, we claim that for $1 \leq i \leq t$,

$$|X_i - X_{i-1}| \leq 2. \tag{4.3}$$

The reason for this is that whether at time-step i the vertex v_i is joined to vertices v_r or v_s does not affect the degrees at later times of vertices v_u, with u distinct from r and s. The joint distribution of all other degrees is the same in either case. As vertices with a given degree are counted, however the degrees of v_r and v_s change in G_t^m, the result is that (4.3) holds. (For an alternate explanation of (4.3) using so-called "half-edges" see the proof of Theorem 1 in [**34**].) By Theorems 3.18 and 3.19 of Chapter 3, we have that with $\lambda = \sqrt{\log t}$

$$\mathbb{P}(|X_t - X_0| \geq \sqrt{\log t}\sqrt{t}) < 2t^{-\frac{1}{8}} = o(1).$$

In particular, X_t converges to X_0 in probability. □

Bollobás and Riordan [**37**] prove the following non-trivial result (whose proof is omitted) which computes the diameter of G_t^m, verifying that the LCD model generates small world graphs. Hence, the LCD model generates graphs satisfying all three of the properties from Section 4.1, and so it is our first satisfactory model for the web graph.

THEOREM 4.4 ([**37**]). *Fix an integer $m \geq 2$ and a positive real number ε. Then a.a.s. G_t^m satisfies*

$$(1-\varepsilon)\frac{\log t}{\log\log t} \leq \operatorname{diam}(G_t^m) \leq (1+\varepsilon)\frac{\log t}{\log\log t}.$$

As with Theorem 4.2, the result of Theorem 4.4 is independent of m. The case where $m = 1$ is not included as the upper bound does not hold there in general.

4.2.3. Rigorous analysis of a PA model. We describe a PA model $\mathcal{G}(m)$ [**138**] which has a simpler description than the LCD model, without the need for the identification of vertices. In addition, the m edges added at time $t+1$ are added *independently*. In the case $m = 1$, this approach coincides exactly with the LCD model. A model analogous to $\mathcal{G}(m)$ for general m was given in [**139**]. Although the independence is less realistic as a model for web page creation, it makes the mathematical analysis easier (however, the proof of the power law is technical).

The sole parameter of $\mathcal{G}(m)$ is a positive integer m. The initial graph G_0 is a fixed finite, connected graph with t_0 vertices and e_0 edges. To form G_{t+1}, we add a single vertex v_{t+1} to G_t. The vertex v_{t+1} is joined via m edges to vertices $w_{t+1,i}$, where $1 \leq i \leq m$, via preferential attachment. More precisely, the probability that $w_{t+1,i}$ is w equals

$$\frac{\deg_{G_t}(w)}{\sum_{u \in V(G_t)} \deg_{G_t}(u)}$$

independently for each $1 \leq i \leq m$. Note that $|V(G_t)| = t + t_0$ and $|E(G_t)| = mt + e_0$. Hence,

$$\mathbb{P}(w_{t+1,i} = w) = \frac{\deg_{G_t}(w)}{c_t t} \tag{4.4}$$

where $c_t = \frac{2(mt+1)}{t}$. Unlike the LCD model, there are no loops in the graphs G_t, but possibly multiple edges. See Figure 4.4 for an example.

The rest of the section is devoted to a rigorous proof of the power law degree distribution for the PA model $\mathcal{G}(m)$. The proof is long, and requires some care. We follow the proof set out in [**138**]. We note that the proof of power law degree distributions usually consist of two parts.

(1) Derive an asymptotic expression for $\mathbb{E}(N_{k,t})$ via a recurrence relation. Auxiliary lemmas concerning convergence of real sequences usually are required.

(2) Show that $N_{k,t}$ concentrates around $\mathbb{E}(N_{k,t})$. This is accomplished using either martingales (as in the proof of Theorem 4.3 in the previous section), or variance (which will be our approach in this section).

We now rigorously analyze the asymptotic behaviour of the proportion $\frac{N_{k,t}}{t}$ and its expected value $\mathbb{E}\left(\frac{N_{k,t}}{t}\right)$. Note that $|V(G_t)| = t + t_0$, but we

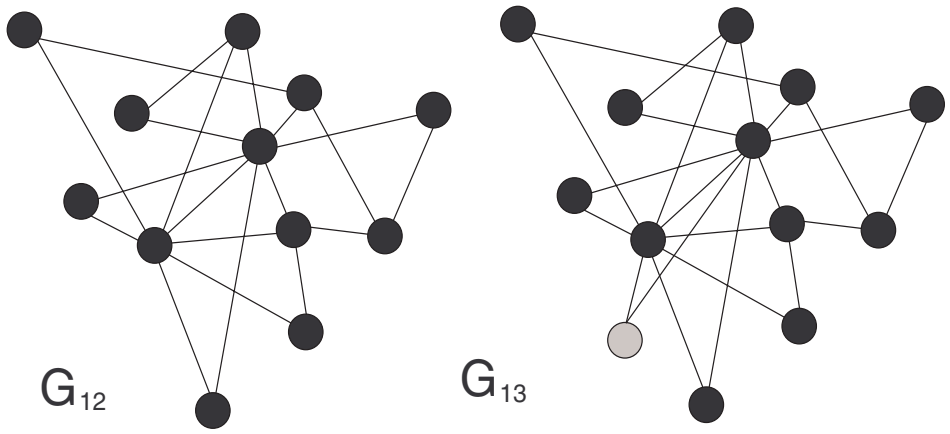

Figure 4.4. The graphs G_{12} and G_{13} in a simulation of $\mathcal{G}(2)$. The new vertex is grey, and is more likely to join to existing vertices with high degree.

divide by t rather than $t + t_0$. Since the results we present are asymptotic, this simplification causes no loss of generality.

We first prove the following lemma from [**138**] about sequences of real numbers, which will be useful for our purpose.

LEMMA 4.5. *For $t \in \mathbb{N}$, let x_t, y_t, η_t, and r_t be real numbers satisfying*

$$x_{t+1} - x_t = \eta_{t+1}(y_t - x_t) + r_{t+1} \tag{4.5}$$

and

(1) $\lim_{t\to\infty} y_t = x$;
(2) *For all t, $\eta_t > 0$, and for all sufficiently large t, $\eta_t < 1$;*
(3) $\sum_{t=1}^{\infty} \eta_t = \infty$;
(4) $\lim_{t\to\infty} \frac{r_t}{\eta_t} = 0$.

Then $\lim_{t\to\infty} x_t = x$.

Proof. By replacing y_t with $y_t + \frac{r_t}{\eta_t}$, in view of item (4) we may assume without loss of generality that $r_t = 0$ for all t. Hence, (4.5) simplifies to

$$x_{t+1} - x_t = \eta_{t+1}(y_t - x_t), \tag{4.6}$$

and so

$$x_{t+1} = x_t(1 - \eta_{t+1}) + \eta_{t+1} y_t. \tag{4.7}$$