Flocks, Herds, and Stories
temporal coherence and the long tail

Mark Bernstein
Eastgate Systems, Inc.
134 Main Street
Watertown MA 02472 USA
+1 617 924 9044
bernstein@eastgate.com

ABSTRACT
New media offer an unprecedented opportunity to revise our literary economy. One crucial anxiety is that we be able to find (and to publish) good work of local or specific importance, since much human knowledge is not popular. Small, low-traffic sites are thus of considerable interest to the health of the Web, though individually these sites possess small economic leverage. The challenge these sites face is increased by the noisiness of web traffic; herds, flocks, and cadres of narrative-driven fans can all increase traffic one day and eliminate it another. For large sites, this poses no problem, but for smaller sites this granularity, combined with the zero lower bound, can have catastrophic consequences both for individual publications and for the overall shape of the Web.

Categories and Subject Descriptors

General Terms
Design, Economics, Human Factors,

Keywords
Web; hypertext; browsing; narrative; flocks; herds

1. INTRODUCTION
New media offer an unprecedented opportunity to revise and reengineer our literary economy[6]. One crucial anxiety is that we be able to learn about good work of local or specialized importance. A great deal of useful and indeed essential knowledge is, at any moment, of pressing interest to a only few people. This does not mean that herpetology, nuclear engineering, classical archaeology, first novels, or malaria are unimportant. The long tail is not a swamp of failure, but simply an acknowledgment that we cannot be deeply engaged with every important and interesting topic.

Attention and traffic are the center of the Web economy. Advertisements convert visits to cash, and cash is readily convertible to visits through advertising and promotion. Unmediated advertising, arbitrarily broadcast to vast numbers readers in the hope of generating a positive return, rewards mere popularity[2], and so values celebrity over achievement, plausibility over accuracy, impact over benefit. Increasingly, search engines and related systems for locating and recommending confront elaborate scams aimed at distracting or misleading readers in the hope that they will, in despair or by accident, click on remunerative ad.

Web Science 2011, June 14-17, 2011, Koblenz, Germany.

The economy of Web traffic is shaped and mediated by links between sites, and also by connections among readers. Those connected readers include individuals, their assistants and their automated agents, institutions such as curated collections, magazines, and indexes, and more autonomous agents such as search engine bots. All of these readers, people and machines alike, discover Web resources by following explicit links, by following implicit or intensive links, or through reference to the sites in other media.

This work begins from hypertext's historic aspiration to shape a superior reading experience for the needs of a disparate audience, and from a simple observation: fluctuations in Web traffic are often quite large – larger than we might anticipate from the simplest models of browsing behavior. Plausible modifications to this simple model readily explain the fluctuations, and also coincide with what we expect and observe actual Web readers to do, and with our own Web behavior. All these models share a common characteristic, that Web readers are not in fact independent entities but instead often form coherent groups moving through the Web together.

This observed coherence, in turn, suggests a variety of interesting possibilities for linking behavior and for the long-term health of the Web – especially the health of the myriad specialized, low-traffic sites collectively known as "the long tail."

2. POISSON NOISE
While our understanding of Web traffic very incomplete, Web traffic is noisier than simple models predict. New Web-borne creations move overnight from obscurity to fame [13], and the power of the Power Law seems unable to save once-popular properties from decay. What leads readers to arrive en masse and then, days or weeks later, to vanish?

Some sources of change are easily understood. Any topic or site, for example, might expect to undergo some long-term *secular* trend as the utility of its information changes, as Web usage grows, and indeed as the population grows. A variety of *cyclical*

changes are easily explained, too, arising from the time of day, the day of the week, and seasons of the year. Exceptional events might also exert large impacts: a power failure might disable the server for a period, reducing traffic. A news event might make one site's information uninteresting while another, formerly obscure, finds itself with information of pressing interest to millions.

It is tempting to ascribe all fluctuations in traffic to the myriad events, known and unknown, that might influence it. Stock market reports, for example, seem always able to explain the day's or the hour's aggregate moves. The Web strategist, too, is eager to see in fluctuations a justification for their revisions and additions to the site. Rationalizations are difficult to test, but they provide a narrative, causal explanation in the face of complex and contradictory forces.

Yet, even in the absence of event-driven disturbance, we would expect some fluctuation in Web traffic. Imagine a Web entirely without links, bookmarks, or search engines. Readers wishing to view a site would type its URL. For simplicity, we further assume that readers never discuss Web sites or recommend sites to other readers, nor do they have favorite web sites that they visit regularly. Even in this extremely simple model, we would not expect traffic to be perfectly consistent. In some intervals, we expect more traffic, in other intervals, less. These fluctuations can be ascribed to two kinds of sources:

- inherent fluctuations arising from external influences, such as news
- statistical fluctuations expected even in the absence of such external influence

In particular, some fluctuations must arise because web traffic is discrete; the number of visits in any interval is necessarily integral. The audience, made up of discrete individuals, is never entirely fluid.

This *Poisson* or *shot noise* is particularly significant for low-traffic sites. Suppose, for example, that we set out to visit an obscure café outside a little-known village because a friend had highly recommended it. We know that it has an open kitchen and three tables, and is open on weekends when the old lady who does the cooking is not visiting her granddaughter. When we arrive, we are not astonished to find the place empty and the cook happy to see us, nor would we be astonished to find three or four other diners there and the cook terribly busy. If we were visiting a large and fashionable nightclub where we would expect hundreds of patrons, and it would be quite surprising to find only two or three others. And if we had gone to see a football match at Amsterdam Arena but saw only two or three other people in the stands, we would be quite certain that some external influence was at work.

More formally, these rare, random, and uncorrelated events obey the Poisson distribution. If, in a given time interval, the expected number of visits is n, the standard deviation of the number of visits is $\sqrt{n}$. If a specialized weblog expects one visitor an hour, then fluctuations of the order of a visitor an hour may be expected routinely, but a sudden influx of – say- 100 visitors would demand an external explanation. A site that expects 10,000 visitors, on the other hand, would routinely experience fluctuations on the order of 100 visitors from hour to hour. But where the low-traffic site's fluctuations of ±1 visitor/hour represent of 100% of the expected traffic, the 100-visitor fluctuations of the busier site represent a noise level of merely 1%.

In practice, observed fluctuations in Web traffic often exceed Poisson noise. Though fluctuations might all be ascribed to external events, several plausible nonlinear phenomena could easily serve to amplify both random and event-driven fluctuations. Though reader behavior is famously complicated [4], simple models can explain the burstiness and temporal coherence of Web readers.

3. HERDS

For many years, my employer has built hypertext tools and published hypertexts. In the early days of the Web, our Web traffic statistics in the Fall and Winter were dominated by (relatively) huge, sudden spikes created by college literature classes. The Web at this time was still new and unusual; a single instructor might introduce a classroom or laboratory full of students to the Web, send them to one of our resource pages, and the sudden influx of hundreds of hits would swamp demand from the rest of the world.

This is a classic example of a herd [9], an ensemble of independent actors adhering to the guidance of one or more leaders. Herding is often observed in social animals, and is a frequent (though implicit) metaphor in Web marketing [7] [2] and in popular imagination of the influence of the Web. Journalists, opinion leaders, and branded aggregators serve as herd leaders, as do collaborative rankings from social sites like Facebook or Reddit. Markov browsing might then induce traffic anomalies at agent sites apparently unrelated to the original target.

Figure 1: A herd arrives in November. Note that the expected Poisson noise for 12K visitors/month is about 100 visitors/month

Strategies for exploiting herds suggest identification and co-optation of prominent influencers and high traffic sites. Cultivation of inbound links increases the likelihood that a herd leader will wander by, followed by the rest of the herd, and hence great effort may be justified in pleasing (or gaining the attention) of strategically-chosen leaders [3]. Outbound links, in turn, are most profitably expended on allied properties that can be relied upon (or required) to cross promote.

An important consequence of herds is that herds increase the granularity, and therefore the Poisson noise, of web traffic. Readers taking part in a herd travel not alone, but rather in groups of thousands or more. If the typical Web reader participates in a herd of m readers, the expected Poisson noise increases with $\sqrt{m}$.

4. FLOCKS

Though cattle and horses travel in herds, the aggregate behavior of many animals depends not on leadership but on comparatively simple and local rules. The "v-formation" flights of geese, or the elaborate coordination of a school of fish, are achieved without leadership; complex group behavior emerges from rules in which each animal attends to its immediate neighbors [9].

Email correspondence, casual conversation, social networks, and dinner conversations readily give rise to flocking on the Web. Although neighbors may not follow each suggestion, and though no individual may share information with many neighbors, short-range information exchange leads to coherent large-scale behavior. Rather than an entire classroom appearing *en bloc* one morning, we may observe unexpected influxes of interest from specialized audiences: suddenly, a product of general interest is sought out by European documentary filmmakers or far-Eastern urban sociologists, simply because these groups talk amongst themselves.

Flocks resemble herds, but offer no leader to coopt. A marketer who seeks to attract a flock needs, in principle, to attract any member, in the hope that the flock will eventually follow[1]. One common approach has been to offer thousands or millions of pieces of focused content – photographs, videos, deviant artwork, discussions of programming questions – in the hope that some will please the flock, just as gardeners cultivate a range of plants in order to attract migratory birds.

5. NARRATIVE

Herds and flocks create *coherence* among readers, increasing the magnitude of apparently-random traffic fluctuations while presenting opportunities to attract large aggregates of readers. The narrative impulse has a similar effect.

The search for narratological closure presents a powerful force for narrative coherence. Presented with a cause, we want to foresee and then to see the effect. Observing a change, people perceive intentional action; seeing action, people want to know how things turned out [5] [10]. Web conventions that provide implicit links to resolve narratives range from Jennicam's daily ritual to quest blogs for books, theses, and diets, from Facebook updates to Twittered revolutions. All reinforce coherence.

The role of narrative in weblogs and in political sites is crucial. As a story unfolds over time, individuals return in order to learn what has happened. Their interest, in turn, often leads them to discuss the narrative with friends and colleagues, and these conversations in turn lead to further narrative engagement. The role of "fandom" and popular media in promoting Web reading has been much discussed, but too often the crucial role of narrative has been submerged in questions of surface design, in extraneous entertainment, or in advertising.

Events and stunts create prodigious, short-lived narrative coherence [13]. Again, narrative coherence increases the granularity of Web traffic; in place of a fluid of myriad individual readers, we have a loosely-coupled ensemble of people who communicate with (some) other readers and whose trajectories are loosely correlated with each other and more strongly correlated with their own history.

6. ONLY CONNECT

The noise level of traffic is of inherent interest to all Web publishers, as it influences resource allocation and expense. To test a Web initiative – an ad campaign, a new design, a better application interface – we must overcome the noise level; in this way, the observed level of fluctuation determines the cost of test driven development for the Web.

A greater concern, however, arises for low-traffic sites, including sites of specialized technical, scholarly, or cultural importance. For these sites, traffic fluctuations represent a significant fraction of total traffic. In addition to the problems posed for larger sites, these traffic fluctuations may approach the zero-traffic lower bound.

The secular growth of a web site's expected traffic T

$$\partial T/\partial t = f(T)$$

may be complex and unknown; f(T) need not even be monotonic for large T where the population of readers may become bored with a popular resource. But, to gain readers, the site must be discoverable, either through links or through search engines. Search engines, in turn, discover and weigh sites by following link to the site. A resource that loses its links can thus be forgotten by the Web and, absent extrinsic efforts to procure fresh links, will receive little or no traffic.

More concretely, let us consider an isolated blogosphere – perhaps a specialized social site – in which N writers participate, adding new material from time to time and linking to other writers on the site whose work they like.

- Each author keeps a list of m favorite weblogs that they read every day. This reading list is published with the weblog.
- When visiting a weblog, each reader follows o outbound links found on that weblog's reading list
- When visiting a weblog that is not currently on the reader's own reading list, the reader may – with probability p – add that weblog to his reading list.
- The reading session continues to depth D.
- Reading lists can accommodate at most C items; if additional sites are added to a reading list, the oldest site on the list is dropped.
- Finally, each writer reviews their own referrer logs, examining referrers for possibly-interesting material. Each day, one recent referrer is examined and may be added to the reading list with probability $r < p$. (Alternatively, we may regard r as reflecting the probability of gaining a link through email, advertising, or exerting some other influence over the weblog writer.)

[1] This can operate through cross promotion within a site's content. *World Of Warcraft,* for example, offers an in-game vanity pet to people who collect Mountain Dew product codes. Flock attraction does require a deeper understanding of what might attract members of a flock; the same promotion might have been less effective with orange juice.

For example, if we have 50 weblogs in our blogosphere, and each reading list can contain 10 weblogs, a typical snapshot might resemble Figure 2. Some sites receive a more traffic and are listed on more reading lists. They continue to receive traffic, and the sites to which these popular sites link also benefit from their popularity. Only eight sites appear on no reading lists and so receive no traffic from the other participants. Even these low-traffic sites, however, are not doomed by their unpopularity; eventually, another writer will follow their referral link and add them to a reading list and they can then rejoin the discourse community.

Shifting to a smaller blogroll of three, rather than ten, items transforms the picture entirely. Now, almost all traffic is concentrated at a few popular sites. Because these popular sites appear in almost every site's reading list, moreover, the traffic-less sites face a much greater obstacle in gaining traction. First, each site is competing with many more sites for the rare opportunity to be cited from referral logs. More significantly, though, since almost all reading lists link to the same popular sites, a site that does manage to reappear on some reading list has only a brief opportunity to gain traction before it is displaced by a link to one of the familiar, popular sites.

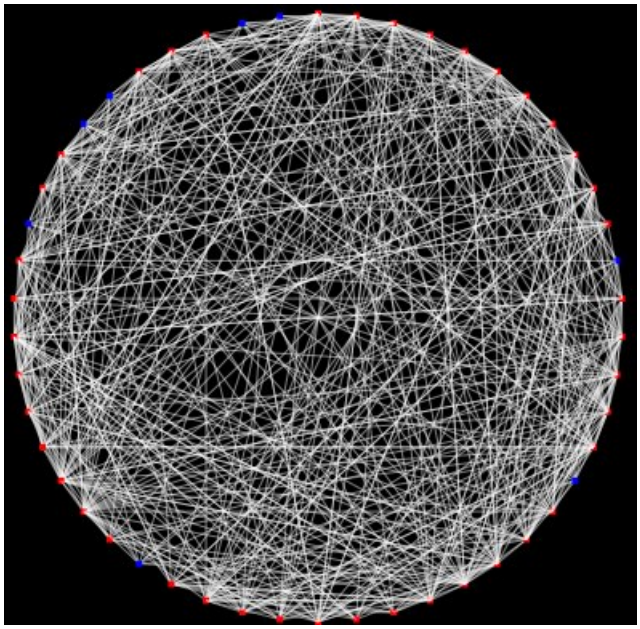

Figure 2: Blogroll linking among 50 weblogs, each with a reading list of 10, after approximately 1 year. (D=4,p=.5, r=.1). 42 sites receive some traffic.

If the probability *r* of gaining a popular blogger's attention is small, the situation is even more dire. A popular weblog can, in this model, endure fluctuations with equanimity; bad days are undesirable but are not a disaster. A less popular weblog, on the other hand, might encounter the zero lower bound as a fluctuation leads it to lose its last inbound links. From this, there may be no recovery.

7. WHY IS THE WEB AS IT IS?

Students of Web phenomenon are naturally tempted to regard the Web as part of Nature, that it unfolds as it always has and, more or less, as it ought or must. But the Web is still new, and quite recently it was entirely possible to envision that the World Wide Web would be quite different from the Web we know. As late as 1993, for example, expert consensus held that the cost of persuading a core of writers and publishers to contribute a small library's worth of information to a public docuverse might be a mere trillion dollars; not only was the estimate quantitatively off target, but the sign was wrong.

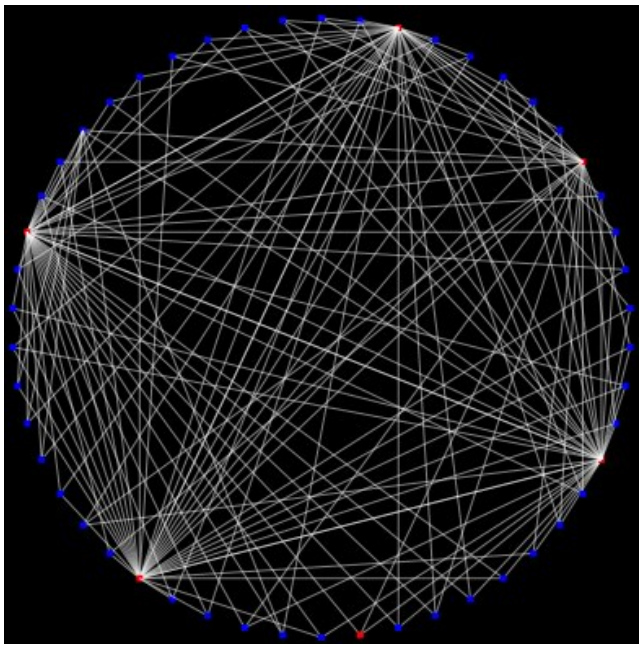

Figure 3: Blogroll linking among 50 weblogs, each with a reading list of 3, after approximately 1 year. (D=4,p=.5, r=.1). Only six notes receive inbound traffic

Indeed, some two years after the Web's popular use began, many expected that its initial efflorescence would give way, concentrating traffic in comparatively few sites.

> *Now, in [January] 1996, I think that web-surfing is dead. Sure, users may check out a few new sites every now and then, just as they may buy a new magazine from the newsstand when they are stranded in O'Hare. But to continue the magazine analogy, most users will probably spend the majority of their time with a small number of websites that meet their requirements with respect to quality and content. Hotlists can only grow so big (especially with the lousy user interfaces for bookmark management in current webbrowsers), so* ***only a few websites will be graced with substantial numbers of repeat visitors****.* – Jakob Nielsen [12]

Indeed, this passage envisions precisely the scenario encountered in Figure 3, where constraints on linking lead to concentration of traffic in a few powerful sites. A year later, Nielsen thought his prediction had been fulfilled.

> *My qualitative data from user interviews and usability studies does tend to support the death of surfing: most users stick to a small number of favorite websites and only go elsewhere when they have specific tasks to accomplish. Most people don't have time to check out*

cool sites just for the sake of looking around on the Web. [11]

Though people did not spend time merely checking out "cool sites," the predicted consolidation of browsing did not take place, and million of sites received, and continue to receive, a regular stream of repeat visitors.

Why did the Nielsen catastrophe not (yet) occur? We may identify several speculative contributors to the preservation of the long tail. These include

- Search engines, which provide authoritative – or at least plausible – starting points for Web readers seeking information on a specific topic, and which prefer a better-targeted, more authoritative site to one that is part of a busy domain.
- Link networks, formal and informal, which effectively spread attention among allied and sympathetic sites.
- Greater, though still incomplete, understanding of the rhetorical utility of links – especially in the role of recommending useful resources as a way to win return visits [14] [1].
- The role of narrative in compelling attention from day to day amongst cadres of readers and writers.

8. RESERVATIONS

The model of reading behavior proposed in sections 6 is simplistic to the point of caricature, and yet contains numerous parameters that cannot readily be derived from empirical measurements. Nor can we hope to derive these measures by fitting simulation to observed patterns of traffic, for the observed traffic is at once too noisy and based on too small a sample of actual Web traffic to allow confident estimation. Sampling total web traffic, moreover, is no longer an option, since so much non-Web information, such as streaming movies and email reading, now travels over http, as does the Web of Data, produced and consumed by computational processes without immediate involvement of a human reader [8].

Nevertheless, the outline of the results appear to be clear. Reader attention and traffic on the Web are not fluid; herding, flocking, and narrative habit all influence traffic and, especially, its variance. The granularity of attention necessarily means that many useful and meaningful sites that would normally expect to receive low or moderate traffic will, from time to time, encounter the zero-traffic lower bound. Finally, it appears to me that any plausible model of Web behavior that depends on links for resource discovery and traffic generation will, in some scenarios, suffer the Nielsen catastrophe in which occasional fluctuations condemn low-traffic sites to oblivion.

Further, most plausible models suggest that the transition from a fairly stable ecology of large and small sites to the Nielsen catastrophe is highly nonlinear. The study of breaksets for random graphs and of percolation through two and three dimensional lattices alike suggest that the transition will be abrupt. A sponge trapped in a pipe can absorb substantial amounts of grit without greatly impeding flow, but as it approaches saturation, it rapidly becomes opaque. In the same way, small policy changes could transform a Web that appeared to be operating well.

9. PROTECTING THE TAIL

In 1996, many expected that the Web might consolidate in a few thousand, or perhaps a few dozen, significant sites. That this did not occur should not lead us to conclude that the Nielsen catastrophe could not have taken place, or might not yet still occur.

Two characteristics contribute to radical consolidation of Web attention and attenuation of the long tail of specialized but valuable sites:

- A limited or impoverished link network, impeding diverse linking practices
- Restricted opportunities for low-traffic sites to regain attention

Subtle policy decisions may exert great influence over these variable. For example, small changes in advertising policy may enhance or diminish opportunities for small sites. If advertising can be effectively targeted to specialized audiences and purchased in small blocks, even a very specialized site may usefully advertise itself. This is, in fact, the state of the Web today. If, on the other hand, advertising is sold *en bloc* to a broad audience, as is often the case in today's "mobile web," specialized sites can neither afford nor benefit from advertising. Similarly, network management policies that privilege transport of data from popular sites (or those that pay large fees to better serve large audiences) could make recovery from the Nielsen catastrophe incrementally more difficult.

We should remember, however, that the principal fact of the literary economy has long been, that books are numerous [6], that the cost of creating and marketing a book is low enough that it books can appeal to specialist readers. Our interest in a wide range of topics is not merely a democratic sentiment or nostalgia for the village community. We may seldom need to know how to repair a broken *sauce hollandaise*, how to introduce a British peer to a Shinto priest, whether the Gallipoli campaign was a good idea, or which of Bunuel's films is most rewarding, but when we do require knowledge, our desire is prone to be urgent. We have long kept libraries and universities as storehouses of unusual but important knowledge, and the future of that indispensable enterprise necessarily lies in the long tail.

10. ACKNOWLEDGMENTS

My thanks to Stacey Mason for reading a preliminary version of this paper, and to two anonymous referees for helpful suggestions.

11. REFERENCES

[1] Ammann, R. 2009. Jorn barger, the newspage network and the emergence of the weblog community, Proceedings of the 20th ACM conference on Hypertext and hypermedia. *HT '09*. 279–288.

[2] Armstrong, T. The AOL Way. http://www.businessinsider.com/the-aol-way

[3] Bernstein, M. 2002. Ten Tips For Writing The Living Web. *A List Apart*. 15 August 2002, http://www.alistapart.com/articles/writeliving/

[4] Bernstein, M. 2010. Criticism. *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. 235–244.

[5] Bernstein, M. 2009. On Hypertext Narrative. *ACM Hypertext 2009*.

[6] Bernstein, M. and Greco, D. 2008. Designing A New Media Economy. *Genre*. 41, 3/4, 59-82.

[7] Gibson, W. 2007 *Spook country*. G.P. Putnam's Sons.

[8] Hall, W. 2011. From Hypertext to Linked Data: The Ever Evolving Web. Hypertext 2011.

[9] Levy, S. 1992 *Artificial Life*. Vintage Press.

[10] Mamet, D. 1998 *Three Uses Of The Knife: on the nature and purpose of drama*. Columbia University Press.

[11] Nielsen, J. Predictions from 1997 Revisited. *Alertbox*. January, 1997, http://www.useit.com/alertbox/9601_revisited.html

[12] Nielsen, J. 1996. Relationships On The Web. *Alertbox*. http://www.useit.com/alertbox/9601.html

[13] Scott, T. 2010. Mob: a flash mob gone very wrong. *Ignite 2*. http://www.tomscott.com/mob/

[14] Winer, D. 2005. The Way To Make Money On The Internet Is To Send Them Away. http://scripting.com/2005/12/12.html