The Emergence of Hypertextual Ecology from Individual Decisions

Miles Efron Steven M. Goodreau Vishal Sanwalani

July 23, 2002

Abstract

Current World Wide Web (WWW) search engines employ graph-theoretic methods to improve their performance. By examining the Web's hyperlink structure, these engines identify topical "hubs" and "authorities." While the success of graph-theoretic based search engines (e.g. Google) suggests that "hubs" and "authorities" play a role in information seeking on the WWW, little is known regarding their emergence. Using an agent-based model, this study explores the conditions that give rise to hubs and authorities. We describe a stochastic network model, with the developers of web sites as agents and the hyperlinks among sites as inter-agent network ties. Agents are provided with finite capital with which to build their sites, conduct research on other sites, and maintain their links; their capital is replenished when their site receives hits from the general population. Our research suggests that the rise of hubs and authorities is a robust phenomenon with respect to initial conditions and various parameter weights. However, the emergence of hubs and authorities is complex insofar as which specific sites become strong hubs or authorities is unpredictable.

1 Background

As the World Wide Web (WWW) has grown in size (current estimates are over 4 billion web pages), the task of locating information has increased in difficulty. Search engines address this difficulty by matching user queries with ostensibly relevant documents. Under traditional information retrieval (IR) approaches, search engines employ a text-based model of relevance; that is, these search engines determine the relevance of a website to a user's search based on the site's textual content.

Recently, Kleinberg [**?**] has proposed retrieval models that incorporate hyperlink information (i.e. how sites are connected to each other). Kleinberg argues that information about website quality is latent in the hyperlink structure defined by the neighborhood of topically related sites. When a user submits a query (e.g. the search string "movies"), often he is interested in locating a website which is an "authority" on a given topic. Kleinberg defines a site to be an authority on a subject (i.e. search string) if it is linked to by many "hubs"; hubs are in turn defined as sites that link to many authorities. Kleinberg provides an algorithm for resolving these recursive definitions and calculating hub and authority scores for individual sites. Applying this algorithm to WWW data, Kleinberg found authorities that match an intuitive sense of relevance [1].

This study describes a model of the process by which hubs and authorities emerge for a single topic domain. We develop a simple network model that mimics the process by which web developers invest in their sites, visit other sites to learn from them, and choose whether or not to create links. All agents share the same goal (to develop a popular website) and the same set of rules for achieving that goal.

[1] The popular search engine Google employs a variation of Kleinberg's model to rank its results.

2 Model

We imagine a one-dimensional "topic space" such as the neighborhood of websites about movies. This space is inhabited by a finite number of agents (websites)[2] with a given location in that space, their "topicality". At time *t*, each agent *i* possesses the following attributes:

- *Topicality* Q_{it} : a real-valued measure on the interval (0,1) of how germane the site is to the neighborhood topic
- *Expendable capital* C_{it} : a non-negative integral measure of how much wealth the owner of a site may invest in it
- *Re-investment aggressiveness* R_i : a real-valued measure on the interval (0,1) of the owner's willingness to re-invest her expendable capital into site development. Note that R_i remains constant throughout time.
- *Hits* H_{it} : the number of hits the site received at the end of round *t-1*.

Each agent also possesses vectors of inlinks and outlinks at every time point. The number of these links yields a site's in-degree (I_{it}) and out-degree (O_{it}), respectively. The structure of these links is used to calculate hub scores H_{it} and authority scores A_{it} following Kleinberg's algorithm [**?**].

All site owners have the same goal: to develop a popular website (i.e. receive many hits/visitors). We assume that a site may increase its popularity via three mechanisms:

- By having significant informational content (i.e. a high Q_{it})
- By containing links to other sites (i.e. a high O_{it})
- By being linked to by other sites (i.e. a high I_{it})

We thus define an additional attribute, visibility (V_{it}), which is a weighted sum of Q_{it}, I_{it} and O_{it}.

Each agent possesses knowledge of her own attributes, and is ignorant of the states of all other sites. However, by expending capital on "research," sites may gain imperfect knowledge of other sites' visibility scores.

The costs and earnings associated with the transactions in which agents engage are defined as:

- K_{link}: the amount of money used to follow an out-link from one's own site to another site, and investigate that site's topicality. In all of our runs $K_{link} = 3$.
- $K_{research}$: the amount of money used to conduct research on another site, learn from its content and decide whether to link to it. In all of our runs $K_{search} = 10$.
- K_{hit}: the expected amount of money obtained for each hit a site receives. In all of our runs $K_{hit} = 1$.

At t = 0, there are no links among agents. In our initial run, each agent is assigned an R_i of 0.5, a $C_{i,0}$ of 10,000 and a $T_{i,0}$ of 0.5. V_{it} was calculated as $0.5 * Q_{it} + 0.5 * I_{it} + 0.5 * O_{it}$ These initial values and visibility formula were changed for subsequent runs, as shown in **??**.

Each subsequent time step consists of the following ordered processes:

- The owner of each site decides how much of her available capital to invest into site development, for a total investment $M_{it} = binomial(n = C_{it}, p = \frac{1}{20} R_{it})$
- Each owner uses $O_{it} * K_{link}$ monetary units to examine her existing out-links and determine whether to continue or discontinue those links. The owner of site *i* will decide to discontinue a link to site *j* if either $Q_{jt} < \frac{3}{2} Q_{it}$ or $H_{j,t-1} < \frac{1}{2} H_{i,t-1}$. Different factors were placed in these comparisons since hit count is expected to be a more volatile measure. If the owner does not

[2] Alternately, one could think of the owner of the website as the agent. Since the links are between websites but the decisions about those links are made by the owners, it is perhaps most helpful to imagine a one-to-one relationship between owners and sites and consider them as interchangeable entities, which together comprise the "agent".

have enough capital to visit all of her existing links (i.e. if $O_{it} * K_{link} > M_{ik}$) then she visits as many as she can afford in random order before stopping.

- Each owner uses her remaining investment money $M_{it} - O_{it} * K_{link}$ to explore other websites in order to learn from them and possibly to link to them. Each owner has only imperfect information about the visibility of each other site. For each site pair i and j, a "research score" R_{ijt} is drawn by multiplying V_{jt} by a random number from a uniform (0,1) distribution. This is called i's research score on j. The sites with the highest research scores are thus likely to be, but not guaranteed to be, those with high visibility. Each owner ranks all sites according to her research scores on them. Starting at the top of her list and moving down, she then takes the following steps until running out of investment capital or visiting every site:
  - Agent i visits a site j at cost $K_{research}$.
  - If $Q_{jt} > Q_{it}$, then site i "learns" from (borrows ideas from, steals content from) site j. This is measured by adding a number from the distribution $N(0.05 * (Q_{jt} - Q_{it}), 0.05)$ to Q_{it}. Since the learned amount is drawn from a normal distribution, it is possible for this amount to be negative; that is, site i actually decreases its topicality by adopting ideas from j.
  - If $Q_{jt} > \frac{3}{2}Q_{it}$ and $H_{j,t-1} > \frac{1}{2}H_{i,t-1}$ then i adds a link to j.
- Each website receives hits from the general public. The number of hits received by site i is a product of its visibility V_{it} and a factor drawn from a binomial distribution (n = number of surfers, p = number of surfers per site). For each hit, it receives a single unit of capital with probability K_{hit}.
- We calculate and store the Kleinberg hub and authority score for each site, following the algorithm described in [**?**]

The model parameters for the twelve runs we conducted are shown in Figure **??**.

3 Results

Under all our models a distribution of hub and authority scores qualitatively similar to Figure **??** appeared. Most sites had intermediate scores on both measures, with tails extending towards high hub/low authority and towards low hub/high authority. No site had both a high hub and authority score (relative to other sites), even though Kleinberg's algorithm for calculating these scores allows for this possibility. That is, distinct hubs and authorities emerged from a population of agents who were all following the same set of rules for creating and breaking links. This observed structure emerged and stabilized rapidly, generally within 30 time steps.

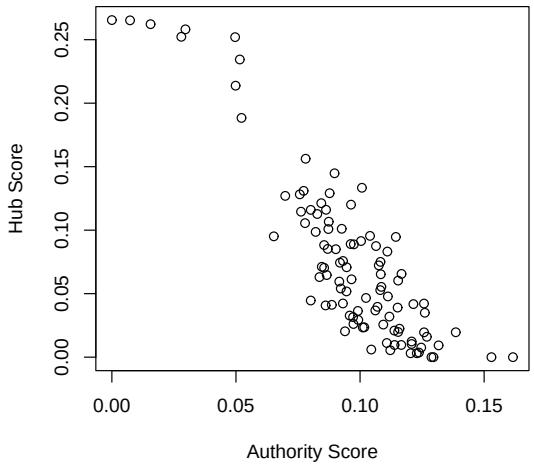

Figure 1: Equilibrium hub and authority scores, Model A

This observation presents an interesting question: Can one predict at the outset which sites will become hubs or authorities? To examine this, we conducted multiple linear least-squares regressions with hub and authority scores as outcome variables and all of the initial conditions as predictors. Linear re-

gression was chosen after examining plots of outcome variables against individual predictors to determine that their relationship could be reasonably approximated linearly. The $t = 0$ values of the predictors were used when these varied among agents; for those cases in which they were initialized equally, the values at $t = 1$ were used instead.

The p-values resulting from these regressions are shown in Figure **??**. R^2 is shown as well, rather than an adjusted R^2, since we are interested in the total predictive power of all initial conditions without any penalty for increasing the number of predictors. One can see that agents' initial attributes do not provide much information about their final positions; that is, the identity of sites that develop into hubs or authorities is generally unpredictable. The regression coefficients are not shown, since the scale of many of these values is arbitrary and not easily interpreted. In models C, E and G, sites were divided into two groups with drastic intergroup differences in topicality, aggressiveness and capital, respectively; yet one could still not predict in which direction members of these two groups would head.

We were also interested in the "economic" rewards of each strategy. Figure **??** shows each agent's final amount of capital plotted against its final hub score. A striking pattern emerges; in this case the two sites with the highest hub score have vastly more capital capital than the other sites, which all score roughly the same. This pattern (with either one or two rich hubs) could be seen in nine of the twelve runs. It persisted even in runs H through K, when the relative contribution of out-degree to visibility was greatly decreased.

Perhaps, even though hub score was not predictable from initial conditions overall, the identity of the one or two wealthy hubs would be. Figure **??** shows initial capital and topicality for Run A, with the wealthiest hub from Figure **??** shown as a solid circle. Examining such plots for all of the runs makes it intuitively obvious that the identity of the wealthy hub(s) cannot be predicted from initial conditions. Once the status of rich hub is achieved, however, it appears to be locked in.

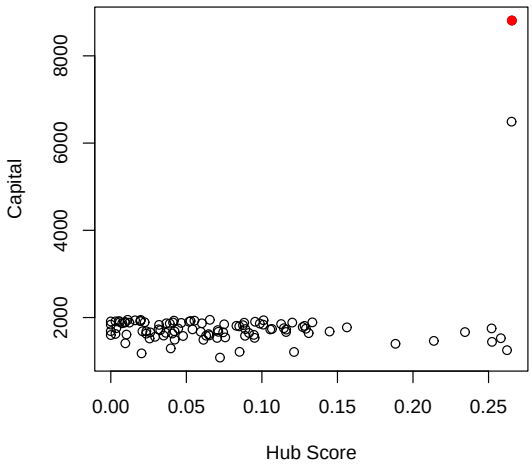

Figure 2: Final capital vs. equilibrium hub score, Model A. The strongest hub is marked by a solid circle both here and in Figure **??**.

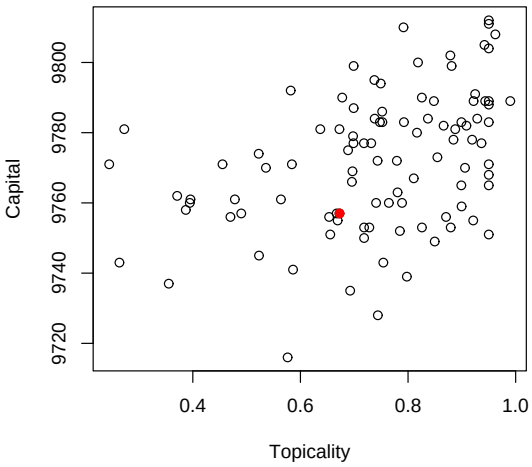

Figure 3: Topicality and capital at $t = 1$ for Run A. The greatest, wealthiest hub at equilibrium is shown as a solid circle near the center of the graph.

4 Discussion

Using a simple model of website development, we discovered the emergence of hubs and authorities to be a robust phenomenon with respect to initial conditions including variation among websites and criteria for website visibility. We also noticed that the hub and authority scores of websites quickly stabilized under all models. Some websites became specialized into certain emergent types (hubs or authorities) despite the fact that all sites were following the same rules. Although the overall structure of the network was robust, the identity of the agents fulfilling the roles of hub and authorities was unpredictable, even when we modified the initial conditions to give a strong advantage to certain sites.

Although extending the results of our model to the World Wide Web is speculative (since our model is static, i.e. the population of websites is fixed, and the WWW is dynamic), the robustness of our model results suggests that even under a changing environment (as the WWW certainly is) the existence of hubs and authorities may be a general phenomenon. Moreover, which sites will develop into hubs or authorities is difficult to predict based solely on initial conditions. The first observation supports the conclusions of Kleinberg. The second matches with our experience of some specific websites; for instance, the website Yahoo began its life following an authority strategy, but has evolved into the most popular hub on the WWW.

References

[1] Kleinberg J 1999. "Authoritative sources in a hyperlinked environment", Journal of the ACM 46(1999). 604-632.

Run	$Q_{i,0}$	A_i	$C_{i,0}$	visibility function coefficients Q_{it}	O_{it}	I_{it}	Agents	Surfers	Time steps
A	all 0.5	all 0.5	all 10000	0.5	0.5	0.5	100	1000	200
B	U(0,1)	all 0.5	all 10000	0.5	0.5	0.5	100	1000	200
C	0.25 for 100 agents, 0.75 for 100 agents	all 0.5	all 10000	0.5	0.5	0.5	100	1000	200
D	all 0.5	U(0,1)	all 10000	0.5	0.5	0.5	100	1000	200
E	all 0.5	0.25 for 100 agents, 0.75 for 100 agents	all 10000	0.5	0.5	0.5	100	1000	200
F	all 0.5	all 0.5	N(10000,2000)	0.5	0.5	0.5	100	1000	200
G	all 0.5	all 0.5	5000 for 100 agents, 15000 for 100 agents	0.5	0.5	0.5	100	1000	200
H	all 0.5	all 0.5	all 10000	0.6	0.3	0.6	100	1000	200
I	all 0.5	all 0.5	all 10000	0.7	0.1	0.7	100	1000	200
J	all 0.5	all 0.5	all 10000	0.1	0.1	1.3	100	1000	200
K	all 0.5	all 0.5	all 10000	1.3	0.1	0.1	100	1000	200
L	all 0.5	all 0.5	all 10000	0.5	0.5	0.5	400	4000	500

Figure 4: Initial parameter values for each of twelve runs.

		Run											
		A	B	C	D	E	F	G	H	I	J	K	L
Authority score	(Intercept)	0.614	0.014 *	0.568	0.371	0.218	0.003 *	0.220	0.714	0.997	0.126	0.435	0.638
	topicality	0.716	0.236	0.365	0.161	0.046 *	0.438	0.437	0.372	0.149	0.232	0.049 *	0.165
	inDegree	0.000 *	0.005 *	0.010 *	0.958	0.710	0.311	0.109	0.547	0.431	0.004 *	0.910	2E-07 *
	outDegree	0.065	0.567	0.127	0.926	0.708	0.028 *	0.362	0.492	0.183	0.066	0.305	0.32
	aggresiveness	---	---	---	0.500	0.175	---	---	---	---	---	---	---
	hits	0.119	0.073	0.462	0.719	0.695	0.943	0.031 *	0.947	0.720	0.710	0.121	0.807
	capital	0.563	0.015 *	0.525	0.386	0.208	0.974	0.136	0.753	0.943	0.146	0.441	0.695
	R^2	0.366	0.458	0.299	0.098	0.111	0.085	0.135	0.035	0.029	0.165	0.071	0.131
Hub score	(Intercept)	0.241	0.09	0.475	0.736	0.588	0.388	0.042 *	0.120	0.706	0.258	0.9244	0.7271
	topicality	0.909	0.0418 *	0.951	0.054	0.023 *	0.295	0.109	0.286	0.678	0.888	0.0152 *	0.001 *
	inDegree	0.079	0.2163	0.426	0.068	0.908	0.346	0.326	0.262	0.618	0.420	0.3419	0.0004 *
	outDegree	0.002 *	0.2915	0.093	0.106	0.089	0.002 *	0.029 *	0.057	0.066	0.002 *	0.9384	0.0492 *
	aggresiveness	---	---	---	0.950	0.676	---	---	---	---	---	---	---
	hits	0.131	0.0098 *	0.807	0.871	0.289	0.411	0.342	0.847	0.729	0.372	0.3284	0.5339
	capital	0.243	0.0785	0.483	0.763	0.602	0.229	0.000 *	0.115	0.722	0.245	0.8719	0.6852
	R^2	0.367	0.483	0.249	0.146	0.214	0.060	0.300	0.140	0.054	0.198	0.077	0.152

Figure 5: P-values and R^2 values from least-squares multiple linear regression; hub and authority scores as dependent variables, initial conditions as predictors