# Diffusion of culture and the "PageRank effect"

Stefano Picascia, Mario Paolucci

LABSS-ISTC-CNR

**Abstract.** We present the Meme-to-Web model - an abstract agent-based model of cultural production and access, aimed at reproducing the dynamics of peer-based content production and retrieval. Agents generate and link cultural artifacts drawing on knowledge both existing and acquired during previous explorations, loosely mimicking the behavior of web surfers. We test three content access strategies: random walking, content searching with information overload, and a mix of the two. Results show a clear leveling effect of PageRank searches on the distribution of visits to artifacts, an effect also documented in related studies. At the same time the mediation performed by the PageRank in the access phase also affects content production and generates very different network shapes that share certain properties with real world web and blog networks.

## 1 Introduction

Ubiquitous personal computing, broadband Internet access and open publishing platforms - the "Read-Write-Web" - have yielded a revolution in the history of culture, at least in two key aspects: (*a*) the symbolic production of our society(es) - our *collective memory* - is today almost entirely stored on the Internet; (*b*) a new paradigm in the production of cultural goods has emerged, one that is based on a wealth of small, independent, mostly non professional producers who disseminate content through the Web. According to some theorists, notably Yochai Benkler [2], such paradigm is bound to replace the hub and spoke model of broadcasting media and cultural industry.

A mostly underestimated effect of the cultural production entirely moving online is that, as Gloria Origgi [13] puts it,

> the Web presents a radical change in the conditions for accessing and recovering cultural memory with the introduction of new devices for managing meta-memory, i.e., the processes for accessing and recovering memory. [...] The way we retrieve information is an epistemic activity which allows us to access through the retrieving filters, how the cultural authorities on a piece of information have classified and ranked it within that corpus

The devices which Origgi refers to are the collaborative content filters, search engines, microbroadcasting applications, reputation systems, and similar tools. Mechanisms that recognize no cultural authorities, but rely on the unattended

interaction of large amounts of faceless, dispersed, heterogeneous users in a bottom-up fashion. Emerged in the first place to provide a shortcut to information in the overloaded web, despite only responding to engineering constraints, these systems have a huge epistemic influence, in that they shape the way we access the increasing share of the cultural production of our society that is developed and distributed via the Web. To start enquiring into the implications of the increasing reliance on these platforms we built a simulation model of a society based on the peer-based production of cultural goods and technology mediation on their distribution. In this work we present the outcomes of a study performed on the most influencing and rooted filtering technology, to some extent embedded into the *www* fabric, Google's PageRank [3].

## 2 The Meme-to-Web Model

The Meme to Web (*MtW*) model implements an artificial society characterized by peer-production and consumption of cultural artifacts, modeled after the insights of [2]. In the *MtW* model a substantial share of the population produces cultural goods ("artifacts") which are made available to the whole population at no charge. An artifact can be thought of as a song, a blog post, a piece of video, a photograph: any reproducible aggregated item of knowledge, such as those produced or transmitted via a computer on the Internet. In our abstract implementation each artifact is no more than a set of memes, that are - in a neglected, but still useful metaphor [5] - basic items of knowledge existing in the form of atomic signs contained in agent's minds and embedded in cultural artifacts[1]. The model, therefore, contains a hypothesis on content creation, namely that information artifacts are:

- created by associating basic idea components (memes) that populate electronic artifacts and agents' minds, both acting as containers. While minds are dynamic containers (the memes contained change in time as a result of agents actions), artifacts - once instantiated - cannot be modified;
- connected to each other *on the base of memetic content*

The circulation of ideas and thought (i.e. of memes) takes place through the distributed production and circulation of the artifacts, which - once created individually - are stored in a separate network of directed links between artifacts.

[1] We take this as a working hypothesis, that is, we are not pretending that these knowledge units really exist - it is the object of a heated debate that we don't plan to enter here. Memes are still to be found in the wild, and they don't have any obvious physical counterpart, nothing even vaguely comparable to the DNA for the genes; still, they are a fascinating hypothesis, the idea of mirroring what we know of the interplay between the genetic code and evolutionary pressure. Skipping the debate, we will just work out the model as if these unit of knowledge existed, hoping to draw an indirect confirmation of our assumption by the plausibility of the network that we are going to grow.

The access to the artifacts in the network is mediated by a set of algorithms modeled after filtering technologies at work on the web.

In the version of the model presented here artifacts are accessed by users in three different ways according to experimental conditions: *RandomWalk*: simply random-walking the graph by following links from an artifact to another; *PageRank*: searching known memes via the mediation of PageRank*; Hybrid*: mixing the 2 strategies above (i.e. searching for a meme and then following the links). In the model, agents search and consume information and the result of hypothesized user activity gives shape to the actual network of artifacts, which reflects again what was inside the minds of publishing agents.

### 2.1 The agent's mind

Agents are built with a simple cognitive infrastructure, consisting in three containers: `{beliefs}` contains memes accepted as true; `{meme-memory}` contains memes encountered during surfing; `{art-memory}` contains artifacts encountered during surfing. We model the belief base as a "meme store". When a new artifact is encountered the memes contained are, at first, stored in a limited memory, and the artifact itself is remembered for a limited time. The set of memes contained in an artifact can be either coherent (in our view: containing at least a meme already believed by the agent), or containing completely new information[2]. In the first case, according to [4] there is a large probability of acceptance of one or more of the other memes contained in the artifact. In the latter case the new memes will have a smaller chance of being remembered and will have a harder time in being incorporated in the belief base. We designed the architecture keeping these simple principles in mind, so if the artifact contains at least one known meme it will be retained in memory according to a probability `sticks-in-meme`, while for completely unknown information the chance of retaining is set to `sticks-in-mem/2`. In this first implementation we have a simpler mechanism for belief manipulation: we use `reiterations` (the times a certain meme has been encountered) as a measure of the probability that the meme will be embedded in the agent's belief base, as in classical cognitive theory [9].

### 2.2 Simulation cycle

The main simulation cycle consists of the six subsequent functions described in detail below. Table 2 shows a list of parameters and their default values.

**Initialization** Agents are created with an initial random number of memes, a Poisson distribution centered on `NMemes` + 1, plus an average `readingCap`, again Poisson distributed in the population, representing the average number of cultural artifacts "consumed" by each user at each time step. A `pct-publishers` fraction of the agents is then attributed with publishing abilities.

[2] In this version we don't have conflicting memes

**Publishing** The publisher produces an artifact and links it to some memes picked from `{beliefs}`. For determining the publishing frequency we adopt the zero crossing blogging model as described in [6]: at each tick, a publisher is in a state represented by an integer with two possible transitions: with equal probability the publisher either adds or subtracts 1 from his current state and publishes an artifact when his state is 0 (i.e., "in the mood for publishing").

**Linking** The newly created artifact is then linked to some of the artifacts stored in `{art-memory}`. The assumption we make is that the linking happens on the basis of content commonality: once an artifact is created, the agent links it to known artifacts with the most similar meme set. At the same time the author links the new artifact with one of her own other artifacts picked from `{creatures}`.

**Exploration** In this phase we implement the three alternative experimental conditions:

- In the *RandomWalk* case the agent picks a random artifact from his memory and follows the outgoing links for `readingCap` steps, reading artifacts.
- In the *PageRank* condition the agent selects a meme from `{beliefs}` and performs a search for artifacts containing it. The artifacts received in response are the top 10 artifacts that embed that meme, ranked according to their `pagerank` value$0 \leq p \leq 1$, computed for every artifact at each tick. A number of such artifacts, no larger than the user's `readingCap`, will be visited by the user.
- The *Hybrid* algorithm stands halfway between the above. The user selects a meme from `{beliefs}` and performs a search for artifacts containing it - just like in the *PageRank* case. The number of artifacts consumed in this case is only `readingCap` / 2; the rest of the exploration is done following links from one of the artifacts visited, like in the *RandomWalk* case. We consider this strategy as the most realistic one, assuming that the browsing habit of real web users consists of the combined exploitation of the searching facilities and of the hyperlink structure of the web.

Memes contained in the artifacts visited will be stored in the user's memory according to the described constraints, should the meme be already there it would gain one `reiteration`.

**Retaining and corruption** Every time a user encounters a certain meme, it gains a `reiteration` point in {`memory`}. As stated above we use reiterations as a measure of the probability that a certain meme will be embedded in the agent's belief base, so every fixed number of time steps users update their belief base integrating memories with most reiterations. Symmetrically, memories with fewest reiterations are removed. The same applies to beliefs stored in {`beliefs`}: they gain `confidence` when encountered often and periodically those with the lowest `confidence` score are dropped.

**Table 1.** Agents and variables

| Agent | Attribute | Description |
|---|---|---|
| user | `ispublisher` | whether a user is also a publisher |
| | `readingCap` | artifacts a user can read at each step |
| | `{creatures}` | list of artifacts produced by a publisher |
| | `{beliefs}` | memes in belief base |
| | `{meme-memory}` | list of memes encountered in surfing |
| | `{art-memory}` | list of artifacts encountered in surfing |
| | `status` | implements randomwalk for publishing |
| artifact | `pagerank` | pagerank value |
| | `pageViews` | hits received |
| | `creator` | the user that produced the artifact |
| | `{memes}` | memes embedded |
| meme | | |

**Table 2.** Global parameters

| Parameter | | value |
|---|---|---|
| `initialUsers` | Initial # of users | 200 |
| `nMemes` | # of memes circulating in the society | 1000 |
| `pct-publisher` | share of users who are also publishers | 0.25 |
| `NArtefacts` | # of artifacts produced at each turn | 1 |
| `AvgReadingCap` | average number of artifacts read each turn | 5 |
| `sticks-in-mem` | probability that a meme is remembered | 0.85 |
| `memes-in-art` | share of a publishers memes in each artifact | 0.20 |
| `conf-loss` | rate of confidence degradation if memes not reiterated | 0.2 |
| `avg-recpr` | chances that a link to an artifact is reciprocated | 0.05 |
| `max-memes-bel` | limits of the belief base | 50 |

## 3 Results

The model was run for a total of 2000 steps with the parameters shown in Table 2 for all the conditions. Content producers are set to a quarter of the population, thus creating a situation of distributed peer production. We present results of simulated content production and access paired with three exploration strategies (*RandomWalking*, *PageRank* and *Mixed*). We examine the dynamics of traffic determined by each condition and the shape and properties of the networks that emerge from each of the three access strategies.

### 3.1 Exploration phase: traffic distribution

**Table 3.** Gini Coefficient of link distribution

| | $g$ (`nMemes`=1000) | $g$ (`nMemes`=5000) |
|---|---|---|
| RandomWalk | 0.94 | 0.96 |
| PageRank | 0.87 | 0.70 |
| Hybrid | 0.89 | 0.73 |

How do the different exploration strategies distribute users attention and time through artifacts? To measure how the different conditions distribute traffic among artifacts we calculated the Gini Coefficient [10] (ranging $0 \leq g \leq 1$ with 0 in case of perfect equal distribution, and 1 viceversa) of the page views in the artifact population. Table 3 resumes the indices in our three test cases, with PageRank giving the less unequal distribution of hits. The result seems very clear: the PageRank helps distributing visits to artifacts in a more egalitarian way, if compared to pure link following. To test the extent to which the searching algorithm mitigates the rich-get-richer phenomenon we ran a simulation with 5000 memes, instead of 1000, representing in such way a wider spectrum of ideas distributed in minds and, consequently, in the artifacts. Interestingly enough, the Gini index of the *RandomWalking* case remains almost unmodified in both conditions, but in *PageRank* and *Hybrid* - the conditions where topicality matters the most - it decreases significantly. In the *RandomWalk* condition, in fact, the paths to artifacts are only dictated by the network structure which presents, as we'll see in the next section, a scale-free distribution of in-links. Therefore in *RandomWalk* the heavvily linked artifacts are those accessed the most, regardless of the memes embedded. The contrary happens in the search engine case, where traffic is driven towards artifacts with high indegree *relatively to a certain meme.* This translates into a more equal distribution of visits among artifacts when search engines are employed, because visits to artifacts are not entirely structure-dependant. The distribution of links among artifacts is extremely unequal in *PageRank* and *Hybrid* also (see further), but here such inequality matters less. Fig.1 renders explicit the dependence of the distributions of visits to

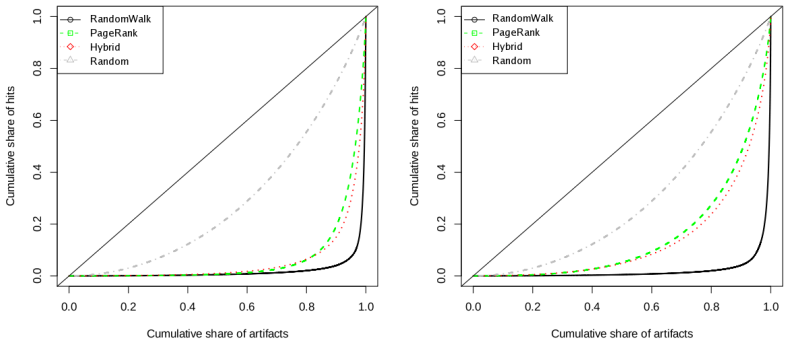


**Fig. 1. Lorenz distribution of traffic.** The line of perfect equality, the diagonal, represents a situation where every artifact has the same amount of visits. The line of perfect inequality coincides with the horizontal and vertical axes, representing a perfectly unequal hits distribution where one artifact has all the hits and everyone else has none. Sandwiched between the two lines is the Lorenz curve. On the left, the Lorenz distribution for a situation with 1000 memes circulating, the right plot shows a condition with 5000 memes. PageRank and Hybrid produce a distribution of hits dependant on the number of memes, and less polarized than RandomWalk.

that of memes. It shows the distributions of our three test cases in the two conditions of 1000 and 5000 memes. We added, for reference, an ideal situation of totally random navigation (grey line), where at each turn agents randomly select `readingCap` artifacts to consume, regardless of links or pagerank[3]. These results seem consistent with a series of works [7,8] that document, both theoretically and empirically, a strong egalitarian effect induced by the usage of search engines on the distribution of visits to websites: a counter intuitive effect, in contrast with the common feeling of a *search engine bias,* or *"PageRank effect"* in favor of already popular sites [12].

### 3.2 Publishing phase: network growth and similarities

We now move on to the other end of the modeled process: the production phase. We examined the network of artifacts emerging from the publishing activity, focusing on the shape of the network and the degree distribution, and compared our artificially-grown networks with two references. The first be the Stanford web graph[4] [11]. This is a fairly extensive collection consisting of a web crawl performed starting from the University of Stanford domain, retrieved in 2002.

[3] Neither this distribution is perfectly linear because of a path dependence (older artifacts are more likely to be visited), but is clearly more equitable than other conditions

[4] Data available on http://snap.stanford.edu/data/web-Stanford.html

The second is a snapshot of the US political blogosphere[5] [1] retrieved in 2004. Table 4 resumes the main features of these networks compared with the simulated ones.

**Table 4.** Network properties

| | Nodes | Edges | Power law exp. |
|---|---|---|---|
| RandomWalk - simulated | 26969 | 144576 | -1.80 |
| PageRank - simulated | 25696 | 167923 | |
| Hybrid - simulated | 25343 | 192020 | |
| Stanford - dataset | 281904 | 2312497 | -1.71 |
| Political blogs - dataset | 1491 | 19090 | |

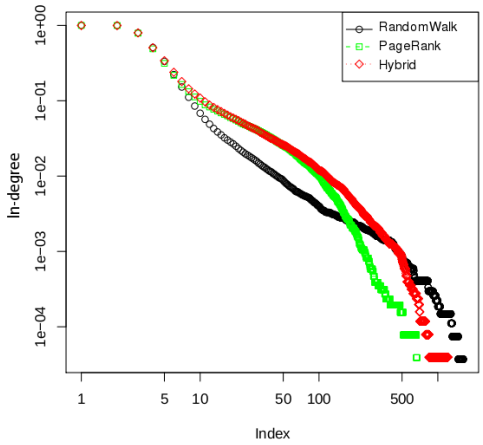


**Fig. 2.** Cumulative node density vs. number of links on a log-log scale for the three simulated networks. Only the *RandomWalk* situation produces a scale-free network

We present the degree distribution of the simulated networks in Fig.2 on a log-log scale. The network built on the basis of random surfing is a scale-free network with the distribution of in-links following a power-law with exponent -1.8. The network springing from meme-based pageranked navigation shows an in-degree distribution which doesn't follow a power law, but is more likely to be fitted with

[5] http://www-personal.umich.edu/~mejn/netdata/

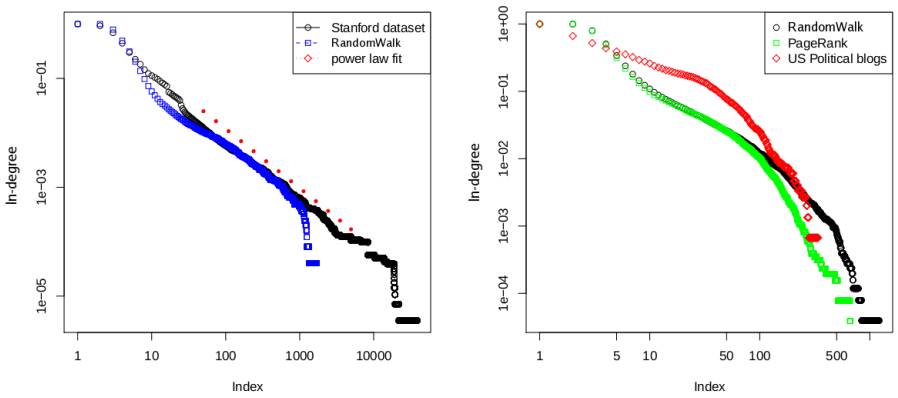


**Fig. 3.** Comparison of simulated networks and two observed distributions - the Stanford network and a collection of political blogs. The Stanford network, a web 1.0 collection, resembles the simulated network generated by the RandomWalk algorithm: they both have a power law zone with very similar slope, -1.8 for RandomWalk, -1.71 for Stanford with a sharp cutoff (left plot). The simulated Hybrid and PageRank algorithms generate a shape that is not a power law, and show a similar trend with the Political blogs network (right plot).

a log-normal curve. In other words, it looks like that the mediation performed by an information filtering technology, such as PageRank, reverberates on the production of content, skewing the distribution of links away from the regularity of the power law that would have emerged. Reading, as detailed above, may result in a modification or the readers beliefs and in turn - for publishing agents - can influence the creation of new artifacts, regarding the level of what memes they contain, and the creation of new links. The blog data set (Fig.3b) shows a certain similarity in the distribution of links with the PageRank and Mixed simulated conditions. It is useful to note that the characteristics shown by the blog data set are said to be frequent in many sub-graphs of the WWW, namely in clusters of sites gathered around a specific topic of interest [14].

## 4 Future work

The MtW model could be a useful tool to explore the possible epistemic outcomes of the wide adoption of wisdom of crowds applications as systems of meta-memory. It needs to be finetuned and refined to include stronger theoretical assumptions. We plan to progressively add other mechanisms that mediate the access to knowledge, such as those commonly associated with Web2.0: collaborative content filters, social bookmarking applications, social networks, micro-broadcasting applications. The goal will be to explore the properties of

each of these technologies in terms of memetic variation, in order to derive hints on the impact that every filtering technology can have on cultural dissemination and gather some hints on how a society based entirely on peer-production may look like.

## References


1. L A Adamic and N Glance. The political blogosphere and the 2004 U.S. election: divided they blog. *In LinkKDD. Adar, E., and Adamic, L. A*, telligence:207–214, 2005.
2. Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, November 2007.
3. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.
4. Cristiano Castelfranchi. Towards a cognitive memetics: Socio-cognitive mechanisms for memes selection and spreading. *Journal of Memetics*, 5(1), March 2001.
5. Richard Dawkins. *The Selfish Gene*. Oxford University Press, USA, 3 edition, May 2006.
6. Christos Faloutsos, Mary McGlohon, Jure Leskovec, and Micaela Götz. Modeling blog dynamics. In *AAAI Conference on Weblogs and Social Media*, 2009.
7. S Fortunato, a Flammini, F Menczer, and a Vespignani. Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences of the United States of America*, 103(34):12684–9, August 2006.
8. Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. The egalitarian effect of search engines, Nov 2005.
9. G. Gigerenzer, U. Hoffrage, and H. Kleinbölting. Probabilistic mental models: a brunswikian theory of confidence. *Psychological review*, 98(4):506–528, October 1991.
10. Mark S. Handcock and Martina Morris. *Relative Distribution Methods in the Social Sciences (Statistics for Social Science and Behavorial Sciences)*. Springer, August 1999.
11. Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *CoRR*, abs/0810.1355, 2008.
12. F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Googlearchy or googlocracy? *IEEE Spectrum*, 2006.
13. Gloria Origgi. Designing wisdom through the web. the passion of ranking. In *Collective Wisdom*. Cambridge University Press, 2009.
14. D M Pennock, G W Flake, S Lawrence, E J Glover, and C L Giles. Winners don't take all: Characterizing the competition for links on the web. *Proc Natl Acad Sci U S A*, 99(8):5207–5211, Apr 2002.