Diversity of Attention and Symmetry of Media : A Free Culture Research Agenda*

Philippe Aigrain†

September 6, 2009

Résumé

Advocates of a free culture framework have stressed [9] its potential benefits for cultural diversity, for a rich sphere of public expression and for replacing the separation between producers and receptors of works by a continuum of capabilities. These claims have been substantiated by many examples or by heuristic reasoning. In the past years, researchers have started to go a step further by defining indicators that can be rigourously modelled and empirically measured to compare cultural frameworks. This was the case in particular for the distribution of attention given to works in each media and for the degree of symmetry between production and reception of contents. This essay reports on the state of modelling, empirical analysis and understanding of these issues. It proposes to further validate claims that both formal free culture (based on voluntary sharing) and de facto free culture (based on unauthorized P2P sharing) are a powerful source of diversity of attention to works. It outlines a possible research agenda for future free culture research encompassing the comparison between access infrastructures (for instance between BitTorrent and P2P using peering servers instead of tracker sites) and the empirical study of the symmetry of media with a great number of contributors.

1 Background and indicators

1.1 The fundamental equation

Cultural or expressive diversity is multifacetted. It can be analyzed as diversity of sources, as diversity of works, as diversity of access or attention to works, or as symmetry of media (balance between creation and reception).

In a universe where hundreds of millions of individuals or small groups can originate works that are theoretically accessible by all, the diversity of sources can

*This essay was prepared for the Free Culture Research workshop held at the Berkman Center of Harvard University on 23 October 2009.

†Author's affiliation: Sopinspace, Society for Public Information Spaces, 4, passage de la Main d'Or, F-75011 Paris, France. Author's contact email: philippe.aigrain@sopinspace.com

be taken for granted. However, this remains virtual if in practice, people access only a limited number of works. Measuring the diversity of works themselves calls for a judgment on their nature or similarity whose objectiveness seems out of reach. Thus, studying the diversity of attention to works has been a predominant approach for evaluating cultural diversity.

Bibliometrists have long studied the distribution of access to books in a library, and noticed that it seemed to follow a power law, more precisely a Zipf's law[1]. The level of access for each work can be modelled as:

$$z(n) = \frac{\mu}{h_N(a) n^a} \tag{1}$$

where n is the rank of the work (by decreasing popularity), $z(n)$ is the level of access to the work of rank n, N is the total number of works (here books in the library), μ is the mean number of accesses to works, a is the fundamental parameter of the law, and $h_N(a)$is the N^{th} harmonic number:

$$h_N(a) = \sum_{n=1}^{N} \frac{1}{n^a}$$

The Zipf's law-based formula (1) proposed above can be considered as the fundamental equation for all studies on diversity of attention and symmetry of media. As we will see, diversity of attention can be studied by evaluating the best fitting Zipf's law parameter (a) for various situations. In contrast, the study of symmetry of media calls for considering both the average rate of access to works μ and the distribution characterized by (a).

1.2 From science to folklore

The bibliometrists soon remarked that in observed distributions the best fitting Zipf's law parameter was close to 1, which in simpler terms meant that the 20% most popular works accounted for a little more than 80% of accesses. Similar distributions were observed in other contexts and the Zipf's law with parameter close to 1 soon became part of folklore. In 2002-2003, researchers [1, 13] claimed that the Internet was not different and that access to sites or blogs also followed a Zipf's law with parameter close to 1. These studies used incoming links as proxies for access which raises some doubt on the validity of their claims. Sites such as Alexa that use information on navigation by their subscribers to rank sites also report rates of access that seem to correspond to a Zipf's law with parameter 1. It would be useful to know if this is based on real measures or simply the result of applying a standard Zipf's law formula. The impact of the sampling (Alexa subscribers) is unknown. The distribution of attention between sites in the global Web deserves to be further studied to confirm or infirm the parameter close to 1 hypothesis.

[1]George Kingsley Zipf was a linguist studying the statistics of occurences of words in languages. He formulated the law that bears his name in 1935. It was soon applied to other domains: income distribution, population of cities and more prominently, access to works in libraries. The power law type of distribution is reported to have been noticed earlier by J.B. Estoup (see the *Zipf's law* entry in the English Wikipedia).

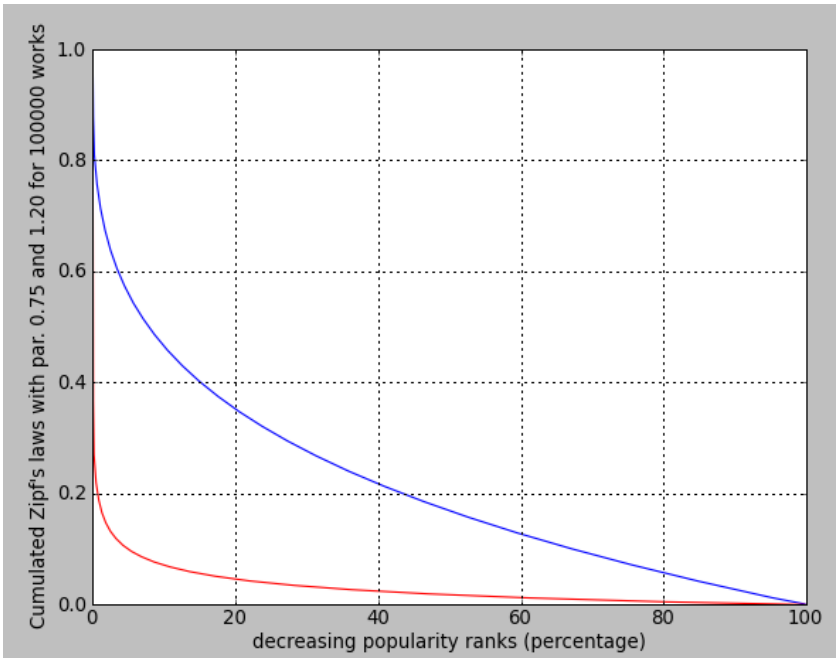

FIG. 1 – Zipf's laws with moderately different parameters can lead to extremely different diversity of attention. The effect is clearly visible when considering cumulated distributions that are classically use to study inequality. Here, in the bottom curve ($a = 1.2$), the 80% less popular works receive less than 5% of attention, while in the above curve ($a = 0.75$), they receive around 35% of attention. Figure from [4].

1.3 The rediscovery of diversity stakes

Analysts of cultural or expressive diversity are interested in access to contents within one media or form of expression rather than in the overall Web. In 2004, Chris Anderson published his famous Long Tail article [7] where he forecasted that the Internet would lead to more diversity of attention, without providing precise modelling of how much. This article was expanded into a book in 2006 [8]. The same year, other researchers [10] provided the first evidence of an increased diversity of sales in the internet channel for products also marketed in physical channels. However, the increase was moderate: in a universe of 20000 products, it would correspond to a Zipf's law of parameter 0.877 instead of 0.935. In contrast, I published in June 2006 results [3] demonstrating that the best fitting Zipf's laws for the distribution of attention to works in various situations have widely varying parameters, ranging from $0,5$ to $1,3$. This range corresponds to sharply contrasted diversity of attention with for instance the 5% most popular works accounting for 22% of accesses in the first case (observed for a music sharing community) and 94% in the latter (record publishing by the major companies)[2]. In recent years, the widely varying diversity of attention was confirmed even in studies of researchers aiming at challenging the long tail theory such as [12][3]. The substantiation of affirmations by empirical observation was until recent work still weak, in particular regarding P2P file sharing. The recent work and agenda in section 2 aims at rooting our knowledge in more solid evidence.

As a further element of motivation, it should be noted than measuring the diversity of attention is not just useful for cultural diversity studies. It is also an essential parameter for the design of measures of usage when collective licensing is used to allow free and legal P2P file sharing. Whether the measure is obtained by monitoring traffic, by mobilizing panels of voluntaries transmtting anonymous data on their usage or by other schemes, a critical design parameter is the minimum threshold of usage that should be measured with reasonable precision. This threshold is highly dependent on the diversity of attention. See the chapter 9 of [4] for details.

1.4 The uncharted continent of media symmetry

Media symmetry is much harder to study than diversity of attention. Actually, its study started in response to criticism against the prospect of a many-to-many information world. Critics started making fun of the millions of blogs, stressing that if everybody writes, nobody will be left for reading[4]. Evidence that many blogs have only very few readers was presented as proof of a dead-end. However,

[2]The small apparent inconsistency between this figure and the number quoted in the caption of figure 1 is due to differences in the size of universes. Zipf's laws can not be perfectly normalized across universes of different size. One should use caution when using Zipf's laws parameters to compare diversity of attention between differently-sized universes.

[3]This article is unclear on many aspects (for instance which peer-to-peer network or protocol was studied). It compares P2P diversity to an extremely diverse theoretical distribution (with parameter of Zipf's law as low as $0,4$) to conclude that the P2P diversity is not as high. Nonetheless it recognizes that their unstated P2P distribution exhibits a significantly higher diversity of access than a (also unstated) legal downloading platform.

[4]See [6] for a review of critics and an endorsement of "expressivism".

this apparent evidence hides a much more complex issue. To start addressing it, one has to ask for each media:

- How many readers (resp. listeners, viewers, etc.) does a piece have in average in a world where all or most are authors? In other terms what is the value of the parameter μ in the fundamental equation?
- Alternately, how much time do people spend reading (resp. listening, viewing, etc.) and how much authoring?

To my knowledge, the study of these questions is still tentative. It may appear that measuring the readership/authorship ratio is not more difficult than measuring the diversity of attention. However, measures that can obtained on the access side (for instance readership for each blog entry in a blog platform) must be complemented by individual-centered surveys if one wants to draw real media symmetry knowledge. Access or visit time measures do not differentiate between time spent editing one's own pieces and time spent reading or downloading[5]. In addition, it can be that people write on one platform and read on others.

2 Diversity of attention: results, challenges and agenda

In [3], I developed a first set of results:

- A method for estimating the best fitting Zipf's law parameter from exhaustive data on access to works in a given universe. This method is superior to classical maximum likelyhood methods. It uses a biased χ^2estimator to derive an initial estimate and a dynamic search algorithm to refine it. When more precision is needed, a dichotomic search can be used[6].
- A demonstration that in most concrete cases the observed distribution is not a Zipf's law in the strict sense[7] but that for all practical purposes of studying diversity of attention, comparison using best fitting Zipf's law are adequate. In particular, when only limited information is available such as "x% of titles generate y% of sales or access", deriving the best fitting Zipf's law parameter from this information provides a reasonable approximation of the general distribution.
- An application of this approach to a limited set of examples for which either full access data or sufficient partial information was accessible,

[5]Researchers such as Sandra Albertolli [5] have used the ratio of the number of unique visitors to the number of active blogs to turn around this difficulty.

[6]The corresponding free software (revised since 2006) can be downloaded at : `http://paigrain.debatpublic.net/docs/Aigrain-distributedcode-articleFM-v3.zip`.

[7]Tested with a Komogoroff-Smirnoff test using tables generated in [11]. It is not surprising that observed distribution are not following a strict mathematical function, as they are the production of compound effecte, some of which are active on specific parts of the distribution, such as promotiing access to the "100 most popular titles".

demonstrating that the attention diversity varied to an extreme degree for various distribution channels and terms of use of a given media (recorded music).

This work suffered important limits, in particular because it did not address the truly large scale access such as P2P file sharing for a given media or access to millions of blogs. An important breakthrough to address the real large scale diversity occurred when a group of researchers of Univ. Paris 6 / LIP6 [2] captured and anonimized 10 weeks of traffic (9 billions messages involving 90 millions users and 275 millions files) for an eDonkey server in early 2008[8].

When full access data is available, one can use a simple Gini indicator to compare diversity of attention. Its computational complexity is much lower than the estimate of the best fitting Zipf's laws. A slightly biased estimate of the Zipf's law parameter can be derived from the Gini indicator. In collaboration with Raphaël Badin, we are presently analyzing the very large dataset provided by [2]. This work should be able to provide a good estimate for the diversity of attention in true P2P networks. In [4], I suggested that the corresponding Zipf's law parameter could be of the order of 0,75., but this remains to be verified or refuted.

A potential further agenda could include:

- Measuring the diversity of attention in tracker-based P2P such as BitTorrent. Researchers are presently starting to work on these domains, for which data collection should be easier. There is a general belief than diversity of attention should be significantly lower in such networks than on peering servers-based P2P networks, because they tend to concentrate attention on recent active titles.
- Scrutinizing some results presented in [2], where the diversity of attention for a specific P2P dataset is presented[9] (graphically) as corresponding to a Zipf's law of parameter close to 1 when some data tables presented in the same paper seem to indicate that the diversity is higher (and thus the parameter lower).
- Negotiating agreements with some of the large sites media contents sites (Flickr, large blog platforms such as Skyblog, Jamendo, MySpace, YouTube) for obtaining access to anonimized data on access, so as to be able to analyse diversity of attention in this context. One could further compare the diversity of attention when contents are licensed under free culture licenses and when they are just made accessible freely without explicit licensing.

3 Symmetry of media: a tentative research agenda

As mentioned above in section 1.4, there are 2 approaches to media symmetry studies, on centered on the number of "readers" per work or productions, the

[8] The dataset is online at: `http://www-rp.lip6.fr/~latapy/tenweeks/`

[9] No precision is given in this paper on which type of P2P protocol and dataset was studied, but repeated references to the Pirate Bay seem to point to BitTorrent.

second based on the ratio of "reading" time budget to "writing" time budget. In [3], I tentatively connected the 2 approaches by conjecturing that within a given media (in the fine grain sense, f.i. blogs or twits), there is a stable relationship between the access ratio and the time ratio. That is:

$$\frac{n_{receptors}}{n_{producers}} \propto \frac{t_{production}}{t_{reception}}$$

Possible approaches to start deriving solid evidence on symmetry of various media could proceed along the following lines:

- When there is a dominant platform within one media (for instance Twitter), directly monitor or negotiate access to anonimized data premitting to compute the access ratio. It would remain interesting to compare the values obtained to those for minority platforms (such as identi.ca in our example).
- In the general case, use detailed surveys on representative samples of users, to study both the access radio and the time budget ratio. Such studies are inherently expensive and complex, but only them can provide true overall knowledge on media symmetry.
- As mentioned when discussing equation 1, it is not only the average ratios discussed above that matter to understanding the degree of symmetry in one media, but also their distribution across individuals and works or series of works.

References

[1] Lada A. Adamic and Bernardo A. Huberman. Zipf's law and the internet. *Glottometrics*, 3:143–150, 2002. accessible at http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf.

[2] Frederic Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an edonkey server, 2008. http://arxiv.org/abs/0809.3415.

[3] Philippe Aigrain. Attention, diversity and symmetry in a many-to-many information society. *First Monday*, 11(6), June 2006. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1337/1257.

[4] Philippe Aigrain. *Internet & Création : comment reconnaître les échanges sur internet en finançant la création.* InLibroVeritas, 2008.

[5] Sandra Albertolli. 12 millions de lecteurs de blogs en france, on s'en fout, non ?, 20 December 2005. Accessible at http://www.heaven.fr/archives/2005/12/12-millions-de-lecteurs-de-blogs-en-france-on-sen-fout-non/.

[6] Laurence Allard. Express yourself 2.0 - blogs, podcasts, fansubbing, mashups... : de quelques agrégats technoculturels à l'age de l'expressivisme généralisé. *Freescape*, 2005. accessible at http://www.freescape.eu.org//biblio/article.php3?id_article=233.

[7] Chris Anderson. The long tail. *Wired*, October 2004. http://www.wired.com/wired/archive/12.10/tail.html.

[8] Chris Anderson. *The Long Tail, Why the Future of Business is Selling Less of More*. Hyperion, New York, 2006.

[9] Yochai Benkler. *The Wealth of Networks*. Yale University Press, ?? 2006.

[10] Erik Brynjolsson, Jeffrey Yu, and Duncan Simester. Goodbye Pareto principle, Hello Long Tail: The effect of search costs on the concentration of product sales, 2006. http://ssrn.com/abstract=953587.

[11] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B*, 41(2), September 2004. preprint from http://arxiv.org/abs/cond-mat/0402322.

[12] Will Page and Eric Garland. The long tail of p2p. *Economic Insight / PRS for Music*, 14, May 2009. .

[13] Clay Shirky. Power laws, weblogs, and inequality, 2003. http://www.shirky.com/writings/powerlaw_weblog.html.