



PROJECT MUSE®

The Web Illusion: Representation and Reality / L'illusion
du Web: Représentation et réalité

Eric Boutin, Samuel Ravatua-Smith

Canadian Journal of Information and Library Science, Volume 40,
Number 2, June / juin 2016, pp. 152-161 (Article)

Published by University of Toronto Press



➔ For additional information about this article

<https://muse.jhu.edu/article/627000>

The Web Illusion: Representation and Reality

L'illusion du Web: Représentation et réalité

Eric Boutin
University of Toulon
boutin@univ-tln.fr

Samuel Ravatua-Smith
I3M Research Laboratory, University of Toulon
drsamuelravatuasmith@gmail.com

Abstract: We are more inclined to visit and eat at a frequented restaurant than a restaurant that is notably deserted. Work on social proof can be transposed into the field of digital technologies. We give more credit to an active and nourished discussion forum compared to a deserted forum. Given this link between quantity and quality, it is interesting to consider the way in which the user builds a representation of the amount of information flowing through a digital device. Is this assessment faithful to reality, biased, overestimated, or underestimated? To answer these questions, we implemented fieldwork in which we asked users from a convenience sample to assess the number of different contributors in a discussion forum. The responses are analysed, and the results are discussed.

Keywords: social proof, discussion forums, cognitive bias, Zipf's law

Résumé : On s'installera plus volontiers à la table d'un restaurant fréquenté que d'un restaurant vide. Les travaux sur la preuve sociale (voir Cialdini, 2001) peuvent faire l'objet d'une transposition dans le domaine des technologies numériques. On accordera plus de crédit à un forum de discussion actif et nourri qu'à un forum déserté. Compte tenu de ce lien entre quantité et qualité, il est intéressant de s'interroger sur la façon par laquelle l'utilisateur va se forger une représentation de la quantité d'information circulant dans un dispositif numérique. Cette appréciation est-elle fidèle à la réalité, biaisée, surestimée, sous-estimée? Afin de répondre à ces questions, nous mettons en œuvre un travail de terrain dans lequel nous demandons à des internautes issus d'un échantillon de convenance d'évaluer un nombre d'intervenants différents d'un forum de discussion. Les réponses sont analysées et les résultats discutés.

Mots-clés : preuve sociale, forum de discussion, biais cognitifs, lois zipfiennes

Introduction

A variety of digital devices provide a way of interpreting data with logical interactions. For example, when connecting to a discussion forum, we visualize exchanges that occur between message emitters and receptors. When we log into Facebook, we may discover messages left by our friends. When connecting to

Twitter, we produce information and encourage people to follow it. In all of these settings, we are not confined to viewing informational content but are also able to exchange content (which may also develop a viral nature). Information about the emitter and receiver permits an understanding of the dynamics involved in constructing a debate, controversy over a subject, and the dialectical process through which we must pass to advance an idea or build a project. In addition, while reading multiple discussion threads in a popular forum, we forge a representation of the dominant point of view and the opposing perspectives and develop an interpretation of the size of the group that presides over these exchanges. In this study we attempt to analyse the representation that an outsider develops while observing a debate in the digital environment, notably after viewing interactions among various implicated actors in a social network. If we consider a discussion forum, through its discussion threads, how many contributing Internauts are hidden behind the forum?

Work in the field of information retrieval in digital environments has focused on the issue of informational and cognitive overload encountered by users (Spink and Cole 2005; Bawden and Robinson, 2008). Research on cognitive biases in information retrieval has documented the representations that users forge in the virtual environment that are far from reality (Lau and Coiera 2007). In this work we focus on another parameter that is likely to influence user heuristics in information retrieval situations. This parameter is related to the interactive nature of information in digital environments. Several digital devices provide data feeds following an interactive logic. When we connect to a discussion forum, we visualize exchanges between transmitters and receivers. When connecting to Twitter, we forward information to followers. In all these devices, we are given informational content in a context of interaction. Our question addresses *the representation that users forge* of the total number of participants actively involved in a debate *after observing interactions between actors on a Web forum*.

In this work we applied an exploratory research method with the participation of 97 university students who explored a discussion forum in a controlled laboratory setting. After viewing the forum, students estimated the number of transmitters and receivers involved. These estimates are compared to the actual number of contributors, and the results are analysed and discussed.

The relationship between the quantity and quality of information

The research question may seem ad hoc and peripheral, yet these representations are at the heart of recent work in information retrieval and information culture. Studies show that in interfaces such as discussion forums, the amount of available or perceived information is based on interpretive heuristics and beliefs. The social decision scheme model indicates that individual and group bias depends on several factors, including group size (Kerr, MacCoun, and Kramer 1996). For example, a discussion forum with many contributors will be interpreted as more trustworthy than a deserted forum. Collective opinions are formed in a condition known as the “wisdom of the crowd” (Surowiecki 2008). In general

Web logic, the proliferation of independent opinions is likely to reveal, by aggregation, information that has high reliability. This kind of thinking gives great value to initiatives like Wikipedia, where, under certain conditions, the quantity of information creates the quality. The idea that the plurality of information is fundamental to democracy has been defended by Dominique Wolton (1997), who addresses the notion of “*expanding the circle of those who speak*” [emphasis added]. The plurality of information also helps prevent the manipulation of the minority. The problem today is that digital devices have developed the appearance of plurality while in reality only a minority are speaking and acting (Adamic and Huberman 2002, 149). If quantity is a factor in quality, the perceived representation of quantity will have an influence on the user’s perception of information quality. To advance the logic even further, for content producers there are certain benefits in making their Web interfaces *more complex than necessary to create the idea that information is more abundant* than is really the case, thus the risk of developing forum interfaces that do not obey a logic of “simplicity” (Berthoz 2009).

The cognitive-bias approach

This study was inspired by the unexpected results obtained during a previous study (Youssef, Boutin, and Souari 2008) that focused on an informetric analysis of a discussion forum. We observed the users’ tendency to overestimate the number of contributors involved in a forum. In this article we offer an interpretation of the results based on cognitive biases that are developed during the information retrieval process (Alonso and Hua 2005; Boutin 2006). For Richard Heuer and the US Center for the Study of Intelligence (1999, 2), cognitive biases are “predictable mental errors caused by simplified information processing strategies.” We propose an interpretation based on a cognitive bias called the “law of small numbers.” In information retrieval the law of small numbers, as originally proposed by Amos Tversky and Daniel Kahneman (1971), refers to the idea that Internet users base their judgements on insignificant samples and tend to have exaggerated confidence in the validity of conclusions based on small samples. From this perspective, it’s interesting to create an analytical framework based on the concept of cognitive bias in information retrieval as a possible factor explaining the observed behaviours.

Laws of the Web

Web laws are observed patterns of phenomena or behaviours that occur among Internet users (Bhattacharjee 2012, 3). We use Web laws as a rule of thumb, such as the 90–9–1 principle and Zipf’s law pertaining to user participation in Web communities (Carron-Arthur, Cunningham, and Griffiths 2014). In essence, Zipf distributions describe situations where large events are rare and small ones are very common (Adamic 2000). While conducting information research on the Web, users are faced with statistical distribution phenomena that differ from traditional statistical frameworks to which they may be accustomed (Taleb 2008). If we look at the number of interventions on a discussion forum, we notice that a small number of actors make most of the contributions,

while the majority of users are small contributors. Here we acknowledge the presence of a Web law characterized by a Zipf distribution (Huberman 2001, 105), a standard feature in online interactions but one that is relatively unobserved by Internet users. Faced with this kind of statistical distribution, there's no guarantee that Internet users will adopt new heuristics different from those used when they are confronted with a normal distribution of data.

In this experiment we used an exploratory methodology with the participation of 97 university students who were asked to navigate and explore a popular discussion forum. The students were then asked to estimate how many people they thought were involved in the forum based on their observations. Throughout this article we will describe the results of the analysis. We also propose a discussion of the results and develop an explanation for the *illusion* that users tend to experience, in which their *perception* of a Web community's size differs from the actual *reality*.

Experimental method: Analysing user perceptions of the size of a discussion forum

This experiment aims to understand the perception that Internauts have of a set of Web interactions. Internauts were put into a situation in which they explored an educational and family-oriented discussion forum entitled Forumfr.com (<http://www.forumfr.com/f525-education-et-famille.html>).

The participants, 97 undergraduate commercial and marketing students (University of Toulon), were divided into five groups. The experiment took place in a controlled setting, and participants had individual computer stations. The experimenter gave instructions to the participants with the aid of a video projector, which was used to present the discussion forum being evaluated. The various features of the forum were presented. The forum contained questions and responses proposed by various users. Each page of the forum contained a maximum of 20 questions. From the time of its creation, the forum had accumulated a total of 63 pages with 20 questions on each page. All of the questions in the forum were open to responses from the Internet community. In addition, the number of responses figured next to each question and was viewable by the public.

Once these clarifications were made, participants were given the URL of the forum and informed that they had 10 minutes to evaluate the website. Once the time expired, the experimenter distributed a five-question survey. The survey was completed individually at each participant's computer station. The heart of the survey consisted of the first two questions, which asked the students to estimate the total number of people involved in the discussion forum (users who asked or answered questions) and the number of users who had intervened at least twice on the discussion forum (asking or answering questions).

Results

There was a wide dispersion of responses to the first two questions. When we asked the users to estimate the total number of actors who intervened on the

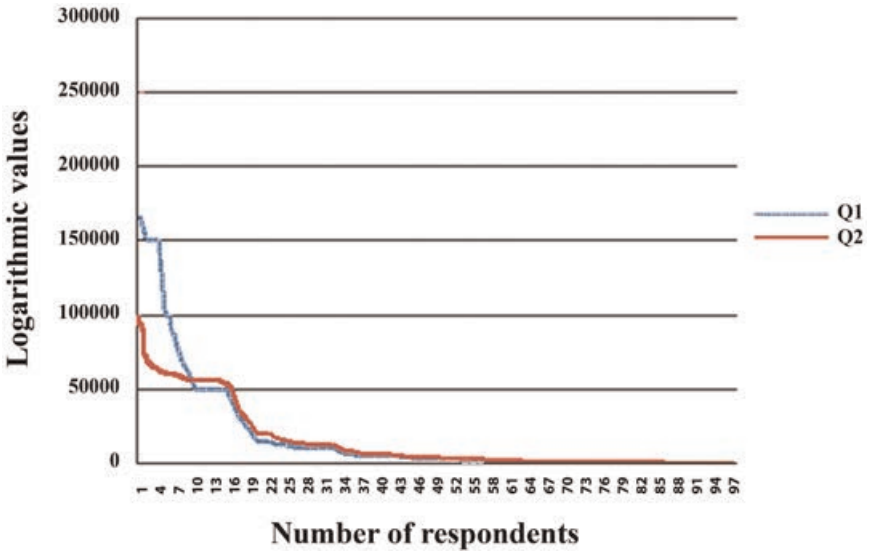


Figure 1: Participant responses to survey questions 1 and 2
Question 1: How many different people were working on this forum by either asking or answering questions? Question 2: How many people intervened *more than once* on this forum by asking or answering questions?

forum, the amplitude of responses ranged from 10 to 166,370 (a ratio of 1:16,000 between the lowest and the highest responses). Responses to question 2 ranged from 5 to 100,000 (a ratio of 1:20,000 between the lowest and the highest responses). We chose to represent the diverse responses with a logarithmic scale graph. The results of both questions are included in figure 1.

Several observations can be made based on this graph. First, we observe that the two representations are almost overlapping. In both cases, the distribution of the observed phenomenon can be approximated by a straight line. The distribution is shown in a semi-logarithmic scale and shows equal numbers of participants provided responses in the following intervals: 100–1,000, 1,000–10,000, 10,000–100,000. This dispersion is surprising and suggests that, in the face of information overload, users do not have benchmarks and develop representations based on their own unique set of heuristics.

In a second phase, we compared these perceptions with reality. On this forum, there has been a flow of 25,294 interventions since its creation. These 25,294 interventions were produced by 3,500 contributors who were involved at least once by asking or answering a question. Of these 3,500 users, 2,000 contributed more than once by asking or answering questions. We considered respondent estimations to be false if they deviated more than 20% from the true value. We considered respondent estimations correct when they were between 2,700 and 4,200 for the first question and between 1,600 and 2,400 for the second question.

Table 1: Cross analysis between gender and responses to question 1

	Male	Female	Total
Q1- Correct	3	4	7
Q1- Overestimated	17	25	42
Q1- Underestimated	15	33	48
Total	35	62	97

Note: Question 1: In your opinion, how many different people were working on this forum by either asking or answering questions?

When comparing the actual results with the respondents’ perceptions, we see that for question 1 (How many different actors have intervened on this forum?), only 7% responded correctly, 49% underestimated the number, and 43% overestimated it. When considering question 2 (How many people intervened more than once on this forum?), we see that 6% of the respondents answered correctly, 62% underestimated the number, and 31% overestimated it.

The results from the two questions show that roughly 7% of respondents provided relatively correct responses that deviated less than 20% from the true number. This percentage is somewhat surprising. It reinforces a general feeling that users create representations without a strong basis in methodological reasoning.

1. The observed results also undermine the principle of “the wisdom of crowds” (Surowiecki 2008). The wisdom of crowds considers that when a phenomenon (in this case the number of people who intervened on a forum) is being estimated, the average responses offered by a large number of unrelated respondents provide a better estimate than a single expert would. In this example, the average estimate provided in response to question 1 was 17,618. This average representation is very far from the reality, with a true value of 3,500. Similarly, the average estimate provided by the 97 respondents for question 2 (How many people intervened more than once on this forum?) was 5,656, still far from the true value, at 2,000.

In addition, we used a chi-square analysis to analyse the variances in accuracy between male and female responses (table 1).

The chi-square analysis shows that the gender of the respondents had no significant impact on the accuracy of the responses. However, it’s important to note that out of the 97 student respondents, 62% were female. Women tended to underestimate, while men had a slight tendency to overestimate the number of participants in the discussion forum. Results from survey questions 3 and 4 indicate that 71% of the respondents found the discussion forum interesting, but only 16% of the respondents would consider contributing to the forum by asking or answering questions (figure 2).

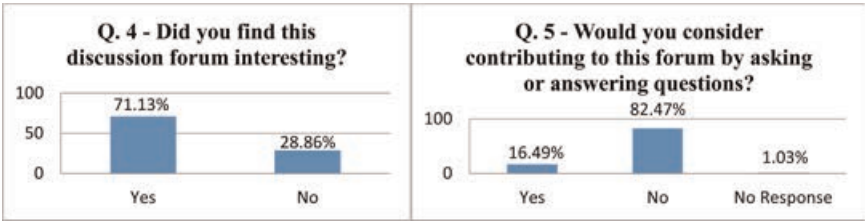


Figure 2: Responses to survey questions 4 and 5
Question 4: Did you find this discussion forum interesting? Question 5: Would you consider contributing to this forum by asking or answering questions?

Discussion

This research highlighted two interesting and unexpected observations: the wide dispersion and a general inability to correctly estimate the number of actors involved in the Web forum. This discussion aims to provide some explanations of the wide dispersion of responses and the general difficulty users had in correctly assessing the number of actors involved in the Web forum.

Evaluating the dispersed nature of the responses

Estimating the number of participants in a discussion forum is difficult because it depends on the richness of interactions among the members of the forum. The difficulty of estimating the size of the community behind the forum may arise from the topology of the network interactions. Some typical examples are presented below:

- In the extreme situation where everyone is related to everyone, n interactions are expressed by the square root of n actors. In this case, we observe a dense community with actors who weave systematic links with each other.
- In another extreme situation, the n links between actors can be expressed by $2n$ individuals. In this extreme example, each actor converses with only one other actor.
- A final example could occur in which two actors alone perform the n interactions by performing n exchanges between themselves.

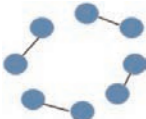


In conclusion, it appears that the number of actors behind a network of interactions ranges between 2 and $2n$, depending on the network topology. These categories are portrayed in table 2.

If we apply this rule to the case in question, keeping in mind that there were 25,294 interventions, the number of actors involved in the forum should be interpreted as being between 2 and 50,588. We observed that 85% of the respondents surveyed gave an answer within this range.

The issue of misperceptions

We observed that Internauts had a distorted view of the number of actors in the discussion forum. It is clear from table 2 that a variety of situations depend

Table 2: Relationship between interaction and actor in connection with the topology of the graph

Number of interactions	Topology: Graph of interactions	Number of actors
n		$2n$
		$SQR(n)$
		2

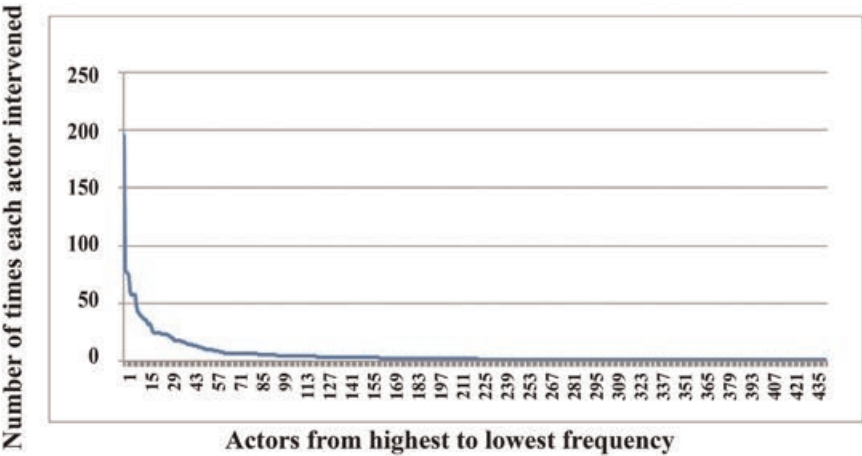


Figure 3: Frequency distribution of actor interventions

on the topology of the interaction graph. In practice, when studying the interaction between actors in digital environments, we observe that the distribution of interactions follows Zipf’s law (Zipf 1949). This means that a small number of actors have many relationships with others and that the majority of actors have only a few relationships. These Zipf distributions constitute a type of Web law typically found in digital environments (Adamic and Huberman 2002; Lafouge and Pouchot 2012). To illustrate this law, we represented the distribution of intervention frequencies that occurred during the first 3,182 interactions. The results are shown in figure 3.

If we organize the actors by the decreasing number of interactions, the product rank frequency is constant. These distributions are hyperbolic and characterized by a long tail. Traditional descriptive indicators characterized by mean and standard deviations are ineffective in describing Zipf's law. We believe that Internauts familiar with "normal" statistical laws are not accustomed to Zipf law. So when the user attempts to estimate the number of contributors on a forum, a sub-set of interventions are considered, and the total number of interventions is inferred. The estimate depends on the sub-set that the user constitutes.

Conclusion

The results show a dispersion of interpretations and a general inability to estimate the size of the discussion forum. The Zipf character of the Web laws observed may partially explain the respondents' difficulty in accurately estimating the number of people involved in the forum. The existence of two polar situations (strongly under- or overestimating) remains to be explained. We must remember the limits of this study performed with a sample group. Our approach was exploratory and merits replication with other sample populations and forums.

This work opens perspectives for future research. It would, for example, be interesting to "dress" a forum with a more or less complex interface to evaluate the impact that interface complexity has on the amount of information perceived by forum visitors. It would also be relevant to examine the link between the quantity and quality of information perceived to identify evidence of other social phenomena in the field of information retrieval.

References

- Adamic, Lada A. 2000. "Zipf, Power-Laws, and Pareto—a Ranking Tutorial." Information Dynamics Lab, HP Labs, Palo Alto, CA. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>. Accessed 23 May 2016.
- Adamic, Lada A., and Bernardo A. Huberman. 2002. "Zipf's Law and the Internet." *Glottometrics* 3 (1): 143–50.
- Alonso, Rafael, and Li Hua. 2005. "Combating Cognitive Biases in Information Retrieval." In *Actes du colloque 2005/International Conference on Intelligence Analysis*. <https://www.e-education.psu.edu/drupal6/files/sgam/Combating%20Cognitive%20Biases%20in%20Information%20Retrieval.pdf>.
- Bawden, David, and Lyn Robinson. 2008. "The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies." *Journal of Information Science* 35 (2): 180–91. <http://dx.doi.org/10.1177/0165551508095781>.
- Berthoz, Alain. 2009. *La simplicité*. Paris: Odile Jacob.
- Bhattacharjee, Anol. 2012. *Social Science Research: Principles, Methods, and Practices*. Tampa: University of South Florida, Scholar Commons. https://eclass.hua.gr/modules/document/file.php/DIT205/Books/Social%20Science%20Research_%20Principles%20Methods%20and%20Practices_Bhattacharjee_2012.pdf.
- Boutin, Eric. 2006. "Biais cognitifs et recherche d'information sur Internet. Quelles perspectives pour les indicateurs de pertinence des moteurs de recherche." In *Actes du colloque VSST*, 1–7. Lille, France: VSST. <http://www.xploorew.com/V SST/Seminaire/06-Lille/B8%20Internet%202/Boutin.pdf>.

- Carron-Arthur, Bradley, John A. Cunningham, and Kathleen M. Griffiths. 2014. "Describing the Distribution of Engagement in an Internet Support Group by Post Frequency: A Comparison of the 90-9-1 Principle and Zipf's Law." *Internet Interventions* 1 (4): 165–68. <http://dx.doi.org/10.1016/j.invent.2014.09.003>.
- Heuer, Richards J., and the US Center for the Study of Intelligence. 1999. "Psychology of Intelligence Analysis." Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.
- Huberman, Bernardo A. 2001. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, MA: MIT Press.
- Kerr, Norbert L., Robert J. MacCoun, and Geoffrey P. Kramer. 1996. "Bias in Judgment: Comparing Individuals and Groups." *Psychological Review* 103 (4): 687–719. <http://dx.doi.org/10.1037/0033-295X.103.4.687>.
- Lafouge, Thierry, and Stéphanie Pouchot. 2012. *Statistiques de l'intellect: Lois puissances inverses en sciences humaines et sociales*. Paris: Publibook.
- Lau, Annie Y. S., and Enrico W. Coiera. 2007. "Do People Experience Cognitive Biases While Searching for Information?" *Journal of the American Medical Informatics Association* 14 (5): 599–608. <http://dx.doi.org/10.1197/jamia.M2411>.
- Spink, Amanda, and Charles Cole, eds. 2005. *New Directions in Cognitive Information Retrieval*. Dordrecht: Springer.
- Surowiecki, James. 2008. *La sagesse des foules*. Paris: Éditions Jean-Claude Lattès.
- Taleb, Nassim Nicholas. N. 2008. "The Fourth Quadrant Problem: A Map of the Limits of Statistics." *An Edge Original Essay*, Edge. October 10. http://homepage.sns.it/hosni/lori/intranet/readings/Edge_the_fourth_quadrant_map_limits_of_statistics.pdf. Accessed 23 May 2016.
- Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (2): 105–10. <http://dx.doi.org/10.1037/h0031322>.
- Wolton, Dominique. 1997. *Penser la communication*. Paris: Flammarion.
- Youssef, Mohamed, Eric Boutin, and Wahida Souari. 2008. "Les outils de dialogue entre les cultures: Cas des forums de discussion." In *Actes du Colloque Ticemed*, 381–390. Sfax, Tunisie.
- Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Appendix

Five-question survey

1. In your opinion, how many different people were working on this forum by either asking or answering questions?
2. In your opinion, how many people intervened *more than once* on this forum by either asking or answering questions?
3. What is your gender? Male or Female
4. Did you find this discussion forum interesting? Yes or No
5. Would you consider contributing to this forum by asking or answering questions? Yes or No