

# A Real-World Spreading Experiment in the Blogosphere

Adrien Friggeri<sup>\*†</sup>, Jean-Philippe Cointet<sup>\*</sup> and Matthieu Latapy<sup>†</sup>

May 16, 2009

## Abstract

We designed an experiment to observe a spreading phenomenon in the blogosphere. This experiment relies on a small applet that participants copy on their own web page. We present the obtained dataset, which we freely provide for study, and conduct basic analysis. We conclude that, in this experiment, the classical assumption that famous blogs act as super spreaders is not always verified.

**Keywords:** complex networks, graphs, spreading, viral marketing, epidemiology, gossip, measurement

## 1 Introduction

Understanding how information spreads among individuals in a social network is a key issue, which received much attention, e.g. [4, 8, 2, 10, 3]. However, this is a challenging task. In particular, precisely observing such real-world phenomena is far from trivial: in most cases, very limited information is available on the spreading process itself. For instance, we do not know in general who got the information from whom and at which time or which other individuals were

in contact with these ones, and the diffusion itself has to be extrapolated from temporal data [7].

We designed a simple web-based experiment, called *happy flu*, aimed at providing data and insight on these issues. It relies on an applet which spreads among web pages. When an individual encounters this applet on a web page he/she visits, then he/she may copy it to his/her own web page, thus spreading it further. This spreading event is recorded, as well as other key information.

We present here this experiment and the data we collected using it. We conduct basic analysis which show that, in this case, there is no correlation between the popularity of a web page and its ability to spread. This is highly counter-intuitive, and in contradiction with most classical assumptions.

This work belongs to the current effort for collecting and analysing real-world spreading data [7, 11]. Its main strength is that the observed phenomena is a pure spreading of information, with no specific assumption on the underlying network and the way users behave. Also, we provide the data freely for study [5], which is an important contribution in itself.

## 2 The experiment

Our experiment relies on a central measurement machine and an applet written in Flash. The applet has a *Spread me* button which produces, for each user pressing it, a personalised copy of the applet with a unique identifier.

---

<sup>\*</sup>CREA – Ecole Polytechnique, CNRS – 32, boulevard Victor – 75015 Paris. TSV – INRA – 65 Boulevard de Brandebourg – 94205 Ivry-sur-Seine Cedex. ISCFIF – 57-59 rue Lhomond – 75005 Paris

<sup>†</sup>LIP6 – CNRS and UPMC – 104 avenue du Président Kennedy – 75016 Paris – France. `Firstname.Lastname@lip6.fr`

Users may paste it on their own web page in order to participate. As a consequence, the new copy of the applet will appear on their own web page, with its *Spread me* button, and the operation may be iterated.

When the *Spread me* button is used, the applet also sends some information to our central measurement machine, in particular which copy of the applet generated the new copy. As a consequence, we record the spreading of the applet among web pages under the form of a spreading tree: we know for each copy of the applet appearing on a web page the other copy from which it was obtained. We also record basic information on each participant, such as the website on which the applet will appear, the participants IP address and his/her country.

In addition, every time the applet is displayed by any user (not necessarily a participant), it sends a message with the user’s IP address to the central measurement machine. We therefore record the number of times each copy of the applet is displayed, as well as the number of distinct IP addresses responsible for this. We store the IP addresses in a secure anonymised way only, in order to preserve privacy.

Once this infrastructure is defined, we still have to give an incentive for individuals to get involved. In order to achieve that, we designed an appealing interface which displays, on each copy of the applet, the spreading tree induced by this copy, measured by the experiment itself. This way, each participant was able to observe, in real-time, his/her own impact and role in the experiment. Moreover, we explained the principle and scientific goals of the experiment, thus raising an interest in it.

Finally, we ran the experiment from July 08, 2008 to September 18, 2008. Five bloggers were first selected among our relatives and were the first and only participants who obtained a copy of the applet from the home web site of the experiment [5]. As we will see

below, after this initialisation step the experiment started to spread rather quickly. After three days, we launched an announcement on the international mailing-list *SOCNET* [1], with the expectation that members of this mailing-list may be interested in the experiment and thus participate to it. After this, we simply observed the spreading until the end of the experiment.

### 3 Obtained dataset and basic observations

During our experiment, a total of 1 051 copies of the applet were generated, of which 492 had more than 1 unique visitor. We considered that the copies of the applet that did not have any visitor have actually not been published.

These 492 active copies of the applet were displayed 481 477 times in total, by 98 200 unique visitors (identified by their IP address).

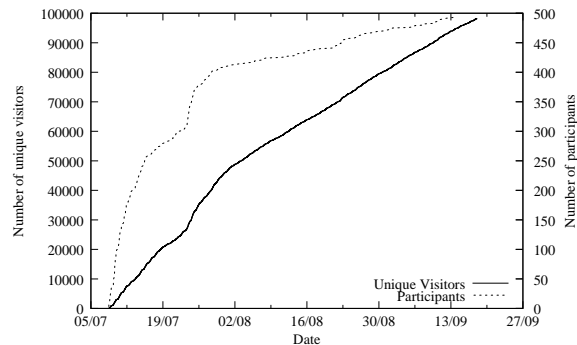


Figure 1: Evolution of the number of participants (right axis) and of visitors (left axis) during the experiment.

The evolution of the number of active participants and of visitors during the experiment is displayed in Figure 1. These plots clearly show two different regimes; we first observed a fast growth in the number of participants during the first three weeks of the experiment and a slower progression thereafter. On July 22, 2008 we made several enhancements to speed up our central measurement machine

which allowed us to serve more applets and hence explains the sudden increase of new active participants at that date.

The obtained dataset is available freely for study on the experiment web page [5] with its full specification, as well as the applet and a video displaying the spreading process during time.

## 4 Super spreaders

One key question for the study of spreading phenomena is the identification of nodes which play an important role in the spreading. In particular, one aims at identifying so-called *super spreaders*, *i.e.* participants who have a strong influence and may induce the participation of many others.

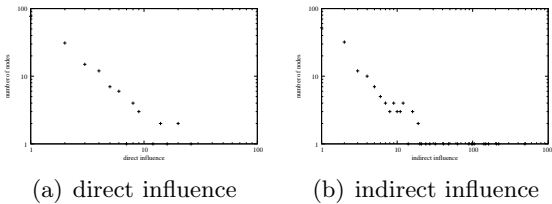


Figure 2: Distributions of direct and indirect influence.

There are several ways to capture this. First, we will call the *direct influence* of a web page  $w$  the number  $d(w)$  of participants directly linked to it, *i.e.* its out degree in the spreading tree. In other words, the direct influence of  $w$  is the number of participants who copied the applet from  $w$ .

Similarly, we will call the *indirect influence* of  $w$  the number  $\bar{d}(w)$  of descendants of  $w$  in the spreading tree, *i.e.* the number of participants who obtained their copy of the applet from  $w$ , or from participants who obtained theirs from  $w$ , and so on.

First notice that, in our experiment, both direct and indirect influences are very heterogeneous (Figure 2), which confirms classical observations of the field and motivates the search for super spreaders.

Notice also that one may imagine scenarios where a participant has a very low direct influence but a very high indirect one. Figure 3 shows that this does not occur here: both quantities are strongly correlated. Moreover, the 6 nodes for which the correlation is the lowest (the ones having a high indirect influence but a relatively low direct one) are nothing but the six initial nodes (the experiment home page and the five blogs we initially used to launch the experiment). They may therefore be considered as a measurement artifact.

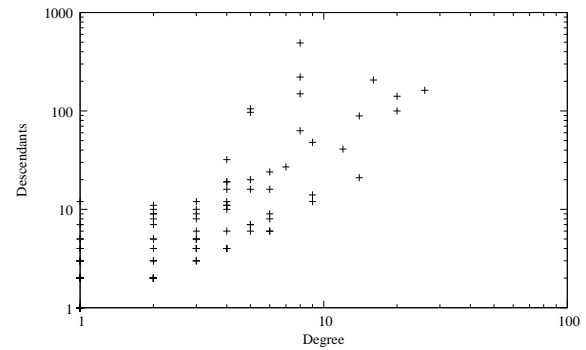


Figure 3: Correlation between direct (horizontal) and indirect (vertical) influence. Both measures are strongly correlated ; the six nodes for which the correlation is the lowest are the six initial nodes.

Finally, as direct and indirect influence are strongly correlated, we will only focus on direct influence here: super spreaders are the participants from which many other participants obtain (directly) their copy of the applet.

A classical assumption in the field is that super spreaders are the web pages which have many visitors, *i.e.* *popular* pages [9, 6]. Indeed, these web pages are supposed to be trusted references for many people, and as they have many visitors they might probably spread the information they publish to many others.

The popularity of a web page may basically be measured as its number of visitors per unit of time. Here, we will capture this by the

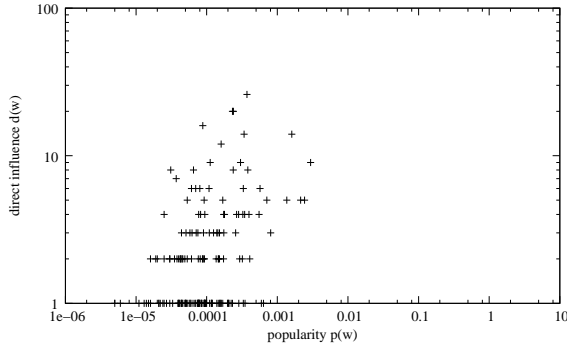


Figure 4: Direct influence  $d(w)$  as a function of popularity  $p(w)$ .

ratio  $p(w)$  between the number of visitors of  $w$  observed during the experiment and the time during which  $w$  was present (*i.e.* the time at which the last hit on  $w$  occurred minus the time at which  $w$  appeared first).

In order to observe the relations between popularity of a web page and its influence, we plot in Figure 4 the influence  $d(w)$  of  $w$  as a function of its popularity  $p(w)$ . This plot shows that there is no web page in our dataset which has a very high popularity but a very low direct influence; conversly, no web page has a very high direct influence and a very low popularity. However, once these extreme situations are eliminated, all other possible cases occur. In particular, some web page with a significant popularity have a high influence, but others have a very low influence; conversly, some web pages with a significant influence have a low popularity. This shows that, in our case, the classical assumptions and intuition stating that influence always is correlated with popularity is false. In particular, the most popular pages are not the ones with the highest influence.

Figure 5 confirms this. It shows that instantaneous influence of our participants (*i.e.* their direct influence divided by the time during which they participate to the experiment) is rather homogeneous: the average rate to which a participant spreads our applet is  $7.61 \times 10^{-7}$  pages per second, the minimum being  $9.154 \times 10^{-8} p.s^{-1}$  and the maximum

imum  $3.54 \times 10^{-5} p.s^{-1}$ . The obtained distribution is far from a power law, the hallmark of heterogeneity expected in such data.

Finally, we conclude that web pages spread our applet at a rather homogeneous rate. In other words, the earlier a participant arrived in the experiment, the higher his/her influence is; popularity has little to do with this.

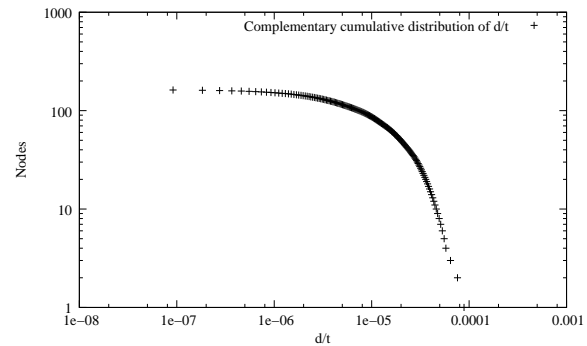


Figure 5: The complementary cumulative distribution function (CCDF) of instantaneous influence shows that the applet is spread at a rather homogeneous rate.

## 5 Conclusion

We designed and conducted a simple web-based experiment aimed at collecting information on how information spread among blogs. This led to the observation of 492 participating web pages during 10 weeks, with 98 200 unique visitors. We recorded the spreading tree and other key information, which we provide freely for study [5].

This dataset is one of the richest ever collected in this field, and makes it possible to observe many interesting phenomena. We illustrate this by computing some simple statistics which show that, in this experiment, the classical assumption that popular web pages are super spreaders is false: the spreading activity of a participant is mostly related to the time at which it joined the experiment, not to its number of visitors.

**Acknowledgements.** We thank Michel

Morvan and Jean-Baptiste Rouquier for their involvement in the conception of this experiment. We also thank *Heaven, Du Marketing Plein Les Doigts* and *GregFromParis* for agreeing to act as starting point of the experiment.

## References

- [1] Socnet@lists.ufl.edu. <http://www.lsoft.com/scripts/wl.exe?SL1=SOCNET&H=LISTS.UFL.EDU>.
- [2] E Adar, L Zhang, LA Adamic, and RM Lukose. Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 2004.
- [3] Lars Backstrom, D Huttenlocher, Jon M Kleinberg, and X Lan. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference*, Jan 2006.
- [4] J Coleman, E Katz, and H Menzel. The diffusion of an innovation among physicians. *Sociometry*, Jan 1957.
- [5] Adrien Friggeri, Matthieu Latapy, and Jean-Philippe Cointet. Happy flu. <http://www.happyflu.com>.
- [6] Kathy E. Gill. How can we measure the influence of the blogosphere? In *Proceedings of the WWW2004 Conf on World Wide Web*, NYC, NY, USA, May 17-22 2004.
- [7] D. Gruhl, R. Guha, David Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th Intl Conf on World Wide Web*, pages 491–501, NYC, NY, USA, May 17-22 2004.
- [8] J Iribarren and E Moro. Information diffusion epidemics in social networks. *eprint arXiv: 0706.0641*, Jan 2007.
- [9] Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. Modeling the spread of influence on the blogosphere. *Proceedings of the 15th International World Wide Web*, page 7, May 2006.
- [10] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. In ACM Press, editor, *Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, 2006.
- [11] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. *arxiv*, physics.soc-ph, Apr 2007.