A Real-World Spreading Experiment in the Blogosphere

Adrien Friggeri[*†]
Jean-Philippe Cointet[*]
Matthieu Latapy[†]
**CREA - CNRS and École Polytechnique*
32, boulevard Victor
75015 Paris
France

[†]*LIP6 - CNRS and UPMC*
Unversité Pierre et Marie Curie
4, place Jussieu
75252 Paris cedex 05
France

An experiment was designed to observe a spreading phenomenon in the blogosphere. This experiment relied on a small applet that participants copied onto their own web page. The dataset obtained which is freely available for study is presented, and a basic analysis is conducted. It is discovered that, despite the classical assumption, in this experiment famous blogs do not necessarily act as super spreaders.

1. Introduction

Understanding how information spreads among individuals in a social network is a key issue that has received much attention; for examples see [1–5]. However, 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 the information providers, and the diffusion has to be extrapolated from temporal data [6]. Another classical approach consists in approximating spreading by citation links [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 that users spread among web pages. When individuals encounter this applet on a web page, they may copy it to their own web page, thus spreading it further. This spreading event is recorded, as well as other key information.

In this paper we present the experiment and data collected using it. We conduct basic analysis which shows 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 analyzing real-world spreading data [6–11]. Its main strength is that the observed phenomena is a pure and true spreading of information, representative of what happens in reality. We moreover provide the data freely for study [12], 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 that produces, for each user pressing it, a personalized copy of the applet with a unique identifier. 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 which copy it was obtained from. We also record basic information on each participant, such as the website on which the applet will appear, the participant's IP address, and 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 and anonymous way 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, their own impact and role in the experiment. Moreover, we explained the principle and scientific goals of the experiment, thus making it more appealing for possible participants.

Finally, we ran the experiment from July 8, 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 website of the experiment [12]. As we show later, after this initialization step the experiment started to spread rather quickly. After three days, we launched an announcement on the international

mailing list *SOCNET* [13], with the expectation that members of this mailing list might be interested in the experiment and thus would participate in it. After the announcement we simply observed the spreading until the end of the experiment.

3. Obtained Dataset and Basic Observations

During our experiment, a total of 1051 copies of the applet were generated, of which 492 had more than one unique visitor. We assumed that the copies of the applet that did not have any visitors were not actually published.

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

The evolution of the number of active participants and visitors during the experiment is displayed in Figure 1. These plots clearly show two different regimes; we 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 that 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 home website of the experiment [12] with its full specification, as well as the applet and a video displaying the spreading process over time.

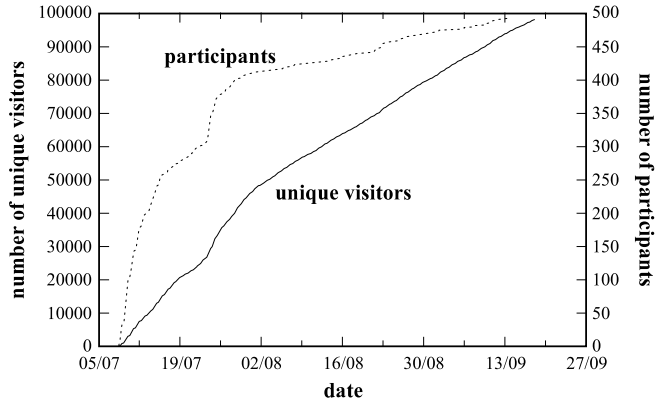
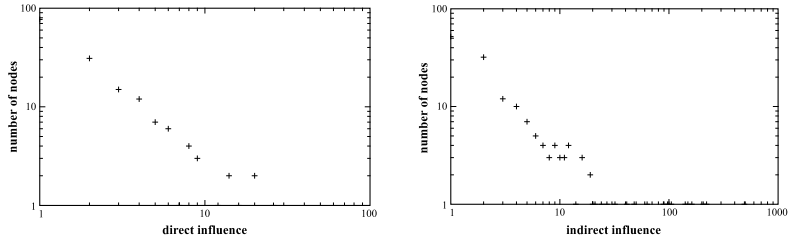


Figure 1. Evolution of the number of participants (right axis) and of visitors (left axis) during the experiment.

4. Super Spreaders

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



(a) (b)

Figure 2. Distributions of (a) direct and (b) indirect influence.

There are several ways to capture a participant's influence. First, we will call the *direct influence* of a web page w the number $d(w)$ of participants directly linked to it, that is, 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 $\overline{d}(w)$ of descendants of w in the spreading tree, that is, 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 note 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.

Note 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 six nodes for which the correlation is the lowest (the nodes with 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 initially used to launch the experiment). They may therefore be considered as a measurement artifact.

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

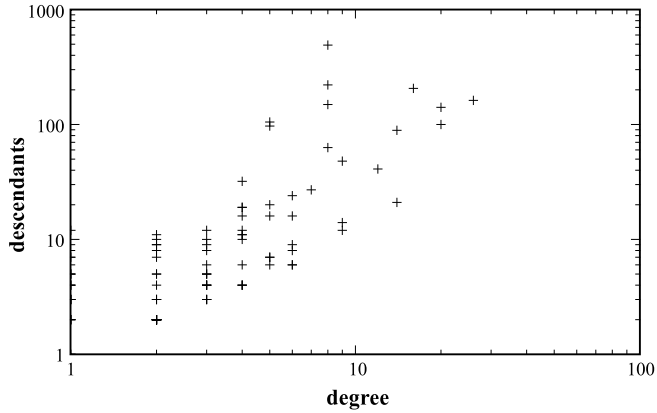


Figure 3. Correlation between direct (horizontal) and indirect (vertical) influence. Both measures are strongly correlated; the six nodes for which the correlation is the lowest are the six initial nodes.

A classical assumption in the field is that super spreaders are the web pages that have many visitors, that is, "popular" pages [11, 14, 15]. 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 capture this by the 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 the popularity of a web page and its influence, in Figure 4 we show a plot of the influence $d(w)$ of w as a function of its popularity $p(w)$. Figure 4 shows that there is no web page in our dataset that has a very high popularity but a very low direct influence; conversely, 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 pages with a significant popularity have a high influence, but others have a very low influence; conversely, 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 is always correlated with popularity is false. In particular, the most popular pages are not the ones with the highest influence.

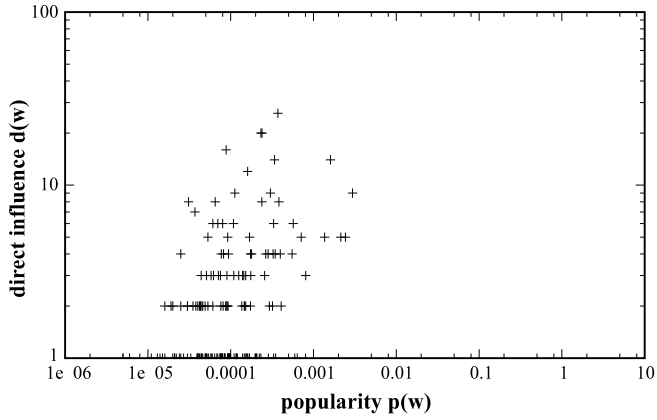


Figure 4. Direct influence $d(w)$ as a function of popularity $p(w)$.

Figure 5 confirms this. It shows that instantaneous influence of our participants (i.e., their direct influence divided by the time during which they participated in the experiment) is rather homogeneous: the average rate to which a participant spreads our applet is 7.61×10^{-7} pages per second ($p.s$), the minimum being $9.154 \times 10^{-8}\ p.s^{-1}$ and the maximum $3.54 \times 10^{-5}\ p.s^{-1}$. The obtained distribution is far from a power law, the hallmark of heterogeneity expected in such data.

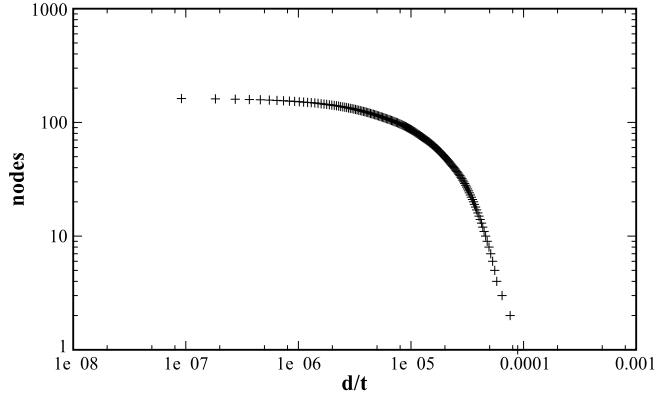


Figure 5. The complementary cumulative distribution function (CCDF) of instantaneous influence shows that the applet is spread at a rather homogeneous rate.

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 their influence; popularity has little to do with this.

5. Conclusion

We designed and conducted a simple web-based experiment aimed at collecting data on how information spreads 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 [12].

This dataset is one of the richest ever collected in this field, and opens the way to the study of 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 the experiment was joined, not to the number of visitors.

Acknowledgments

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 points for the experiment. Finally, we thank Lionel Tabourier for useful comments. This work has been partially supported by the French National Agency of Research (ANR) through grant *Webfluence* ANR-08-SYSC-009 and by the Ville-de-Paris programme Emergence(s) through grant *DiRe*.

References

[1] J. Coleman, E. Katz, and H. Menzel, "The Diffusion of an Innovation among Physicians," *Sociometry*, **20**(4) 1957 pp. 253–270.

[2] J. L. Iribarren and E. Moro, "Impact of Human Activity Patterns on the Dynamics of Information Diffusion," *Physical Review Letters*, **103**(3), 2009. doi:10.1103/PhysRevLett.103.038702.

[3] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. "Implicit Structure and the Dynamics of Blogspace." Presented at the Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference (May 18, 2004) www.blogpulse.com/papers/www2004adar.pdf.

[4] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The Dynamics of Viral Marketing," in *Proceedings of the 7th ACM Conference on Electronic Commerce (EC '06)*, Ann Arbor, MI (J. Feigenbaum, J. Chuang, and D. M. Pennock, eds.), New York: ACM Press, 2006 pp. 228–237. doi:10.1145/1134707.1134732.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group Formation in Large Social Networks: Membership, Growth, and Evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, Philadelphia, PA (T. Eliassi-Rad, L. Ungar, M. Craven, and D. Gunopulos, eds.), New York: ACM Press, 2006 pp. 44–54. doi:10.1145/1150402.1150412.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information Diffusion through Blogspace," in *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*, New York (S. Feldman, M. Uretsky, M. Najork, and C. Wills, eds.), New York: ACM Press, 2004 pp. 345–354. doi:10.1145/988672.988719.

[7] J.-P. Cointet and C. Roth, "Socio-Semantic Dynamics in a Blog Network," in *IEEE International Conference on Social Computing (ICSC '09)*, Vancouver, BC, IEEE, 2009 pp. 114–121. doi:10.1109/CSE.2009.105.

[8] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading Behavior in Large Blog Graphs," *Physics and Society*, 2007. arxiv.org/pdf/0704.2803v1.

[9] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, **329**(5996), 2010 pp. 1194–1197.

[10] M. Cha, A. Mislove, and K. P. Gummadi, "A Measurement-Driven Analysis of Information Propagation in the Flickr Social Network," in *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, Madrid (J. Quemada, G. León, Y. Maarek, and W. Nejdl, eds.), New York: ACM Press, 2009 pp. 721–730. doi:10.1145/1526709.1526806.

[11] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of Influential Spreaders in Complex Networks," *Nature Physics*, **6**, 2010 pp. 888–893. doi:10.1038/nphys1746.

[12] A. Friggeri, M. Latapy, and J.-P. Cointet. "Happy Flu." (2008) www.happyflu.com.

[13] "SOCNET Social Networks Discussion Forum." (Jan 14, 2011) www.lsoft.com/scripts/wl.exe?SL1=SOCNET&H=LISTS.UFL.EDU.

[14] A. Java, P. Kolari, T. Finin, and T. Oates. "Modeling the Spread of Influence on the Blogosphere." Presented at the 15th International World Wide Web Conference, Endinburg, Scotland (May 22–26, 2006) http://ebiquity.umbc.edu/_file_directory_/papers/262.pdf.

[15] K. E. Gill. "How Can We Measure the Influence of the Blogosphere?" Presented at the 13th International World Wide Web Conference, New York (May 17–22, 2004) http://faculty.washington.edu/kegill/pub/www2004_blogosphere_gill.pdf