The Berners-Lee Hypothesis: Power laws and Group Structure in Flickr

Andrea Baldassarri, Alain Barrat, Andrea Cappocci, Harry Halpin, Ulrike Lehner,
Jose Ramasco, Valentin Robu, Dario Taraborelli

Dagstuhl Social Web Communities Working Group
Schloss Dagstuhl, September 2008
Andrea.Baldassarri@roma1.infn.it, alain.barrat@gmail.com, andrea.capocci@gmail.com,
jramasco@isi.it, hhalpin@ibiblio.org, V.Robu@cwi.nl, Ulrike.Lechner@unibw.de,
d.taraborelli@surrey.ac.uk

Abstract. An intriguing hypothesis, first suggested by Tim Berners-Lee, is that the structure of online groups should conform to a power law distribution. We relate this hypothesis to earlier work around the Dunbar Number, which is a supposed limit to the number of social contacts a user can have in a group. As preliminary results, we show that the number of contacts of a typical Flickr user, the number of groups a user belongs to, and the size of Flickr groups all follow power law distributions. Furthermore, we find some unexpected differences in the internal structure of public and private Flickr groups. For further research, we further operationalize the Berners-Lee hypothesis to suppose that users with a group membership distribution that follows a power law will produce more content for social Web systems.

1 Introduction

Despite the advent of the Social Web, where the participation of users is critical, very little is known about the structure of online groups and how these relate to the social contacts, membership in other groups, and even the productivity of users. While intuitively it would seem that a user would join groups that already have many of their social contacts, the possibilities of how group structure, membership, and productivity interact are wide. Is there any correlation between a user joining a group and the number of their social contacts already in the group? Does this depend on group size? A simple hypothesis could be that a user would prefer to join many small groups already containing their friends. However, an alternative hypothesis would be that users prefer to join a few large groups, where users can share their interests with others who they may not yet know.

1.1 The Berners-Lee Hypothesis

Tim Berners-Lee, widely acclaimed as the inventor of the Web, has put forward the hypothesis that "it seems from experience that groups are stable when they have a set of peers, when they have a substructure" so that "neither the set of peers nor the substructure must involve huge numbers, as groups cannot 'scale,' that is, work effectively

Dagstuhl Seminar Proceedings 08391
Social Web Communities
http://drops.dagstuhl.de/opus/volltexte/2008/1789

with a very large number of liaisons with peers, or when composed as a set of a very large number of parts" [3]. In other words, both being a member of only a single large group and just being members of many small groups is less than ideal for users. In fact, Berners-Lee likens a single large group to a "global monoculture" and a large set of small groups to a "set of isolated cults" [3]. Berners-Lee further states that "the compromise between stability and diversity is served by the same amount of structure at all scales," in other words, a "fractal distribution" [3]. His prime example of using this "fractal requirement" to discover "how you fit in to society at large (and at small)" gives a distribution that increases by a constant exponential scale. In other words, Berners-Lee suspects that the ideal structure of groups is given by a power law distribution or, as Berners-Lee puts it, a "Zipf-shaped" distribution [3]. While finding true self-similarity and so fractal structure in groups on the Web may be difficult, finding a power law distribution of group membership is straightforward to detect. While Berners-Lee has "no mathematical theory to demonstrate that this is an optimization of some metric for the resilience of society," the detection of a power law could be a sign of "building an effective society on top of the Web" [3]. The intriguing 'Berners-Lee Hypothesis' would seem to further predict that the most productive users would be members of communities at different scales, scales distributed along a power law distribution.

1.2 Power law distributions and the Dunbar Number

There are intriguing reasons why such a power law distribution may be expected in group size, user membership in groups, and even participation in groups due to a phenomenon called the 'Dunbar Number' [7]. The Dunbar Number is a hypothesized cognitive upper limit to the number of individuals one can form a social relationship with at a given time and this upper limit was estimated by Dunbar to be 150 [2]. However, Dunbar's later work hypothesized that the Dunbar Number of '150' was just one of a series of 'circles of intimacy' in human social relationships [2]. Refining his original hypothesis, Dunbar hypothesized that the number of social contacts people possess follows a vaguely exponential curve, where one in general has 5 intimate friends, followed by 12-15 members in a sympathy group, followed by 150 friends one can maintain, followed by 1500 acquaintances [2]. So, to be precise, there is no single Dunbar Number. Instead, there is a Dunbar distribution, where the infamous Dunbar Number is just an estimate of the maximum number of friends one can actively maintain via contact at a single time, a single point on a distribution of social contacts.

Originally, Dunbar had no human data on social networks, but measured the number of social contacts of primates, such as apes, which were observed by studying primate grooming patterns [7]. Dunbar then extrapolated these results to guess the 'typical number of social contacts' of humans by increasing the observed number from apes in order to compensate for the increase in the size of the human neo-cortex [7]. This was justified since the "cognitive constraint on the size of social networks in those species that live in intensely social groups" may be the result of "the number or volume of neocortical neurons limits an organisms information processing capacity, and hence the number of social relationships that an individual can monitor simultaneously" [9]. There has been research that has found something resembling Dunbar's number in human social networks [9].[1] However, it would be reasonable to believe that just as the advent of language helped increase our cognitive capacities to keep track of our social technologies, we can use social networking Web sites to extend our native cognitive capacities to keep track of social contacts [6]. However, recent research shows that social networking sites report that the online Dunbar Number, reported to be 129 with a deviation of 120, is similar to the off-line Dunbar number [10]. It should be noted that all these results that estimate the Dunbar Number have a huge deviation, likely predicting that social contacts per individual are distributed by a non-Gaussian distribution like a power law distribution. To differentiate this from the Berners-Lee Hypothesis about group size and group membership, we will say that *Dunbar distribution* is the predicted power law distribution of the number of social contacts per user. Although Dunbar's original hypothesis was that there is some uniform maximal number of social contacts due to cognitive constraints is not likely to be correct, his later idea that 'circles of intimacy' follow a power law distribution is of interest and possibly the foundation of the Berners-Lee Hypothesis.

While the precise details of any supposed Dunbar Number are likely wildly varying due to this underlying power law distribution, the Dunbar distribution would help explain the Berners-Lee Hypothesis. If the number of social contacts of an individual was distributed by a power law, this could lead to the hypothesis that the size of groups and user membership (the number of groups an individual is a member of) also follows a power law. If power law distributions are signs of "highly effective" groups, this may be because certain cognitive attention constraints on group participation lead people to belong to multiple groups that follow Berners-Lee's "fractal" scaling, in a manner similar to the limits hypothesized by Dunbar [3]. A productive user would belong to both a few groups that have many members and a 'long tail' of many smaller groups with less members, with social contacts spread throughout both kinds of groups. A user who belonged to many high membership groups of strangers could be overwhelmed by the amount of activity and be unable to meaningfully participate, while a user that belonged only to many small groups that consist of people they already know would become isolated.

1.3 Towards a more testable hypothesis

We will first test for the more basic and well-known Dunbar distribution, and then test the Berners-Lee Hypothesis. We can operationalize both the Dunbar distribution and the Berners-Lee Hypothesis in the following manner:

- The *Dunbar Distribution*: the social contacts of users themselves are spread on a power law distribution, with a few very hyper-connected users and a large 'long tail' of less connected users.
- The *Berners-Lee Hypothesis*: the membership of groups themselves should follow a power law distribution, with a few high membership groups and many low membership groups.

[1] Although it should be noted this was found using the dubious methodology of tracking the distribution of Christmas Cards, which found that a mean social network size of 153.5 with the large deviation of 84.5 [9]. Such large deviations are one sign of a power law distribution.

2 Preliminary results regarding Flickr group structure

What is needed to test the Berners-Lee Hypothesis and the Dunbar distribution is a large data set generated by the activity of real-users. In this paper we have not been able to fully validate all the aspects of the Berners-Lee Hypothesis. However, we have managed to perform an investigation of the structure of Flickr groups that tests a simple form of the Berners-Lee Hypothesis using a large data set crawled from Flickr, and we provide the results below.

First, we should mention the experiments used two large, independently collected Flickr data sets, which were labeled according to their source. One data set, called the "Paris" data set was collected and kindly made available to us by researchers at France Telecom, Paris, while the "Tagora" data set comes from a large European Union project of the same acronym. Unless otherwise noted, our results are based on the "Paris" data set, and we have clearly labeled where this is not the case. Both data sets are very large; for example, the "Paris" data set contains about 1.7 million individual Flickr users, almost 73 thousand different user groups and around 15 million user-user contact relationships.

Note that by "group" in this data set we mean an explicitly joined group with a discrete self-selected number of members. This would be an *explicit* group that vastly differs from an *implicit* group such as those detected by community-detection algorithms. With explicit groups, there is the possibility of joining a group where one literally has no connection to anyone else in the group, while this would not be the case for an implicit group found by a community-detection algorithm.

2.1 The Dunbar Number: Social Contact Size on Flickr

Our first step was to turn our attention to the distribution of number of contacts per user, irrespective of the Flickr groups they belong to. This step tested for the existence of the Dunbar distribution. The results are shown in Fig. 1. The distribution for both the "in-" and "out-" contacts follow overlapping power law distributions, with the characteristic long tail.

From the perspective of Dunbar's theory, however, the "contact" relationship in Flickr must be seen in the broadest (i.e. least intimate) possible sense. In our data set, 8223 users among the 1.69 million had over the 300 contacts, which is similar to the number suggested by Dunbar as an upper bound of the number of friends a person can have based his Christmas card experiment [9]. In fact, many users had contacts ranging in the thousands, with the most connected user having over 12,400 contacts. It is quite likely these people are companies or professional photographers whose number of contacts does not really denote a close social relationship.

2.2 The Berners-Lee Hypothesis: Group Size and User Membership on Flickr

Next, in Fig. 2 we show that the frequency of group sizes, as well as user membership (number of groups a user belongs to) also follow power law distribution. Since group size follows a power law, the Berners-Lee Hypothesis holds in a simple form. As already mentioned, the emergence of power law distributions in group size is not really surprising in this context. Power laws can emerge from any preferential attachment type of phenomena, such as people joining groups which their friends already belong to, as shown earlier [1]. Although previous work does not observe specifically power law behavior in group size, it can be expected that some preferential attachment mechanism could lead to the emergence of a power law distribution in this case.

However, what is more curious for the Berners-Lee Hypothesis is namely that user membership in groups also follows a power law distribution. So, most users belong to a small to medium number of groups, but a few belong to a truly astounding number of groups. This effect is not easily explained by a simple preferential attachment mechanism, and is more likely to be explained by constraints on attention and cognition that limit many users from joining too many groups.

2.3 Comparison of contact and group distributions found in 2 Flickr samples

As Fig. 3 shows the power laws for group size and user membership as shown earlier in Fig. 2, for both the “Paris” and “Tagora” data sets instead of just the “Paris” data set. As both of these are large scale, independently collected data sets, it is not surprising there is a degree of overlap. The difference between the two data sets could be explained by fact that one of the data sets, the “Paris” data set contains more private groups than the “Tagora” data, which is based on public groups. As shown in the next section, the distinction between private and public groups may be important.

2.4 Relationship between group size and internal structure

Finally, the relation between the size of a group (in terms of numbers of members) and its internal structure (measured as the number of internal contact relationships between pairs of users both belonging to the same group) is analyzed in an attempt to test whether or not the Berners-Lee Hypothesis can be built on top of the Dunbar distribution. While specifically power law mechanisms have to our knowledge not specifically been observed in this case, as shown in previous work (e.g. [1]), online users are more likely to join groups that many of their social contacts already participate in.

Fig. 4 shows this result for the “Paris” data set. Here a very interesting and surprising effect is observed. It seems that the distribution of internal structures is diverging, as if there are basically two independent power law distributions present in the data. In one distribution (the one more on the left), the number of users tends to be low and there is a high number of internal social links in the group, and in the second distribution (the one more on the right), there are far fewer social contacts for yet even more members of the group. This could be explained by the fact that some groups are interest-driven, where the members of the groups are strangers yet united by a common interest, versus smaller groups where members know each other.

In Flickr, the most likely explanation for this effect is that it is the result of the private vs. public group distinction. From manual examination of the data set, it seems there are many private groups around interest in sexual and even perhaps pornographic photos. It seems plausible that these private groups consist mostly of a large number of users where the users are not in social contact. In other words, many people view the pictures and they wish to remain unknown to each other. In public groups around less taboo topics, it seems the groups are often smaller but consist of many more people that know each other. As a word of caution, we stress this explanation has not been thoroughly verified, but merely as the most plausible working hypothesis based on a limited manual observation of the data. This hypothesis is left to be confirmed or disproved more thoroughly in future work.

3 Concluding discussion

The Berners-Lee Hypothesis, in its most simple form, seems to hold, although the reasons and implications of this are still areas for future research. In general, group size follows a power law. Furthermore, user membership in groups follows a power law. What would be the clearest test of the Berners-Lee Hypothesis would be to determine if the size of groups a user belongs to follows a power law. Furthermore, the Dunbar Number does not hold, since users seem to have a vast variance of social contacts both above and below 150. However, the Dunbar distribution holds, such that the number of social contacts are also distributed as a power law. These results are not surprising, as many decentralized social and technological systems produce power law distributions, as shown by previous work by some of the authors of this report, especially in the context of collaborative tagging [8, 4, 5]. Further work should be done on creating a carefully grounded model for explaining the generation of these power law distributions that relates them to user behavior and cognitive constraints. The exploration of the precise nature of relationships between these power laws is yet to be done.

While the initial discovery of power law distributions were not surprising, we have found some surprising results when inspecting the intersection of the Berners-Lee Hypothesis and the Dunbar distribution in attempting to predict whether or not group size and user membership in groups can be correlated with the number of social contacts in the group. In particular, there seems to be two distinct distributions, which reflect the disjoint internal structure of public vs. private Flickr groups. A more thorough exploration of this issue is also needed.

Furthermore, one could imagine a number of effects that could quantify Berners-Lee's comments that users that follow a power law distribution in their group membership are highly effective. Is there any correlation between membership in groups and the amount of content generated? Does this depend on group size? A simple hypothesis could be that many small groups would encourage user contributions. However, an alternative hypothesis would be that a few large groups would be more stable and more likely to generate activity. To operationalize this notion, a system displaying a power law distribution of membership in groups could be:

- Demographically more stable over time than a system with a uniform distribution of members per group.
- Likely to have a power law distribution of user activity within each of its groups.
- Likely to produce a higher overall amount of activity over all groups compared to a system with any other group membership distribution

Lastly, the relationship between user productivity and group structure should also be explored, and we have made tentative steps in this direction by at least formulating some testable hypotheses about the stability and productivity of groups and users. We leave testing these hypotheses to be explored in future work.

4 Acknowledgments

We wish to thank the organizers of the Dagstuhl Seminar on Social Web Communities for inviting us to attend and for their support. Furthermore, we also thank the researchers at France Telecom for kindly providing us a large part of the data used in the experiments.

References

1. Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Disc. and Data Mining*, pages 44–54. ACM, 2006.
2. Louise Barrett, Robin Dunbar, and John Lycett. *Human Evolutionary Psychology*. Princeton University Press, Princeton, New Jersey, USA, 2002.
3. Tim Berners-Lee. The fractal nature of the web, 2007. http://www.w3.org/DesignIssues/Fractal.html.
4. Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, 2007.
5. Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. *AI Communications*, 20(4):245–262, 2007.
6. Merlin Donald. *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press, Cambridge, Massachusetts, USA, 1991.
7. R.I.M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.
8. Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proc. of the 16th International World Wide Web Conference (WWW'07)*, pages 211–220. ACM Press, 2007.
9. R.A. Hill and R.I.M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.
10. W. Reader. Are 'friends' eletric? social network sites and the development of intimacy, 2008. Submitted for publication. http://www.reader.statichost.co.uk/FriendsSNS.pdf.

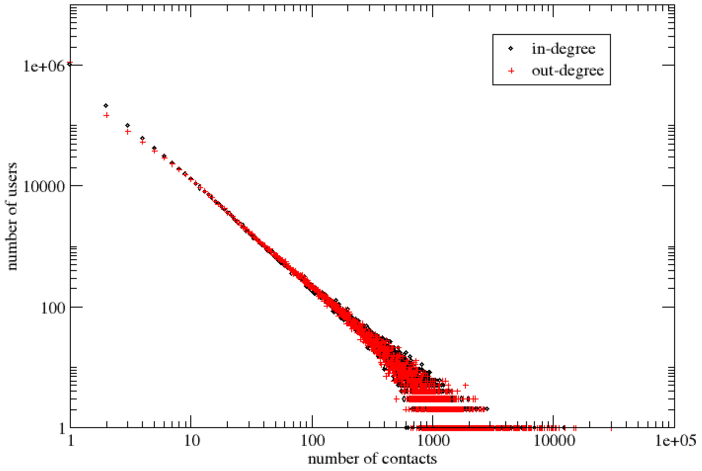

Fig. 1. Distribution of the number of Flickr contacts per user. The horizontal axis shows a number of contacts, while the corresponding point on the vertical axis gives the number of users that have that exact number of contacts. Note that, on Flickr, the “contacts” relationship is not symmetrical. However, the two graphs appear to overlap, showing that, although there is no symmetry constraint imposed by Flickr on contact relations, this symmetry does emerge in practice.

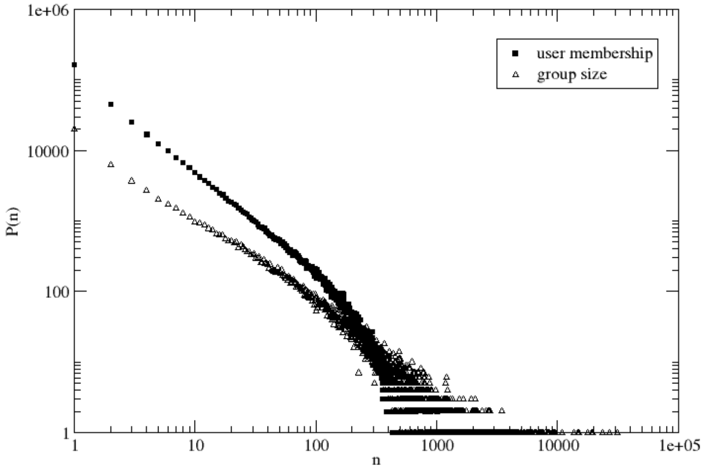

Fig. 2. Distribution of group membership and the size of groups. This is a super-imposed graph: the horizontal axis shows either the size of each group (empty triangles) or the number of groups that a user belongs to (filled squares). The corresponding point on the vertical axis gives the frequency (i.e. number of instances) of the corresponding abscissa point. Hence, there are almost 1 million users belonging to only one group, and over 10000 groups with only one user.

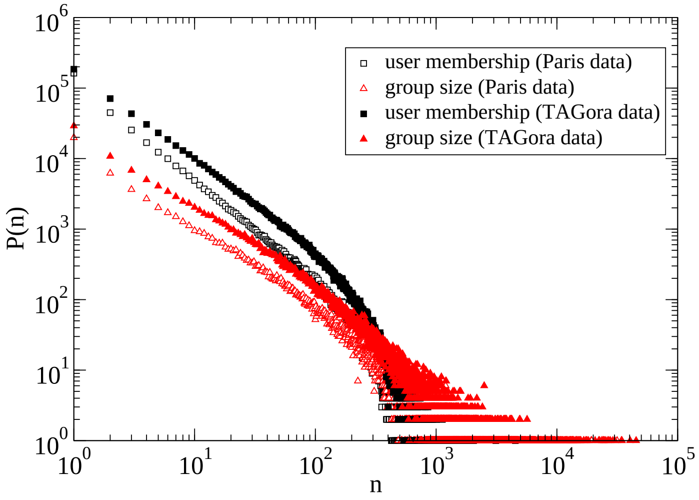

Fig. 3. Distribution of contacts and group membership based on two distinct data sets. Here the meaning of the axis (and actually, one of the graphs) is the same as in Fig. 2. However, the visualization has been extended to two, large-scale and independently collected data sets.

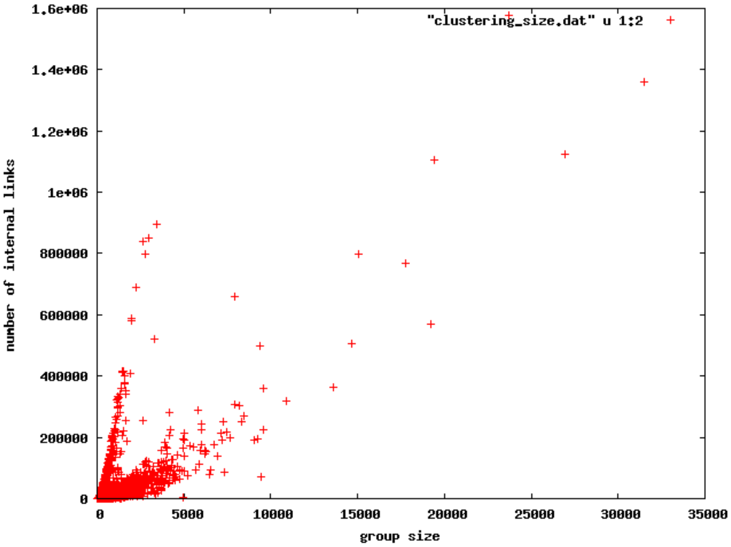

Fig. 4. Relationship between group size (measured as the number of users) and its internal structure (measured as the number of internal links or contact relationships between any pair of users both belonging to the group. Notice the separation into two divergent distributions.