Exploration of the network spun by website users

Ashok Argent-Katwala[1], T.S. Evans[2], and Uli Harder[1]

[1] Department of Computing
Imperial College London
180 Queens Gate
London
SW7 2RH
United Kingdom
{ashok|uh}@doc.ic.ac.uk
[2] Theoretical Physics
Blackett Laboratory
Imperial College London
Prince Consort Road
London
SW7 2BW
United Kingdom
t.evans@ic.ac.uk

Abstract. Using data stretching over more than 5 years for the website `gallery.future-i.com`this paper investigates basic properties of the popularity of pictures and the way users navigate through the website. We find that the rank frequency plot of the downloaded pictures follows a Zipf law. The download rate of individual popular pictures over time resembles that seen of infection rates of diseases. The graph created by successive downloads of pictures shows a power law that does not change over different time periods. We discuss how our findings can be modelled and how they are of importance to website performance.

1 Introduction

In this paper we investigate the website `gallery.future-i.com`. The website has been in existence since April 2001. The website allows registered users to publish pictures in `jpeg` format. Pictures have to be added to an existing or newly created gallery and have a unique name within that gallery. All pictures are accessible to registered and non-registered users. We are in the fortunate position that we have the complete logfiles of the webserver hosting `gallery.future-i.com` as well as information of the growth of the website as the database is run using a `postgres` database. All data will be made available on the web and can be used by other researchers.

This information may give us insight into the ways humans perceive and process such data. While of intrinsic interest in its own right (e.g. see the response to Barabasi's model of queuing [1] or investigations of cultural transmission via simple copying mechanisms [2]) this may also suggest ways we can enhance the performance and useability of similar systems. However unlike sites containing text based information (e.g. ordinary web pages or collections of academic papers) a photo based site such as ours does not currently allow machine based analysis of the items[3] and hence any automatic derivation of their relationships. Rather we must focus on generic methods derived from the way users actually use the site.

In this paper we gauge the popularity of pictures by looking at their download rates. The changes in popularity of a picture has similarities to that found in models of cultural transmission [2] or can be compared against models for the spread of rumours [5] and infectious diseases.

Another aspect we investigate is the way users navigate the website and download pictures. We form a graph by connecting two successively downloaded pictures by the same user. We compare different user types.

Both aspects are important in the context of caching information to speed up the content delivery of the webserver. The first aspect might improve least recently used (LRU) strategies, though theses strategies might already pick up popularity. The second aspect is a more subtle way to predict the next picture a user is likely to download.

2 User navigation

For the years 2002-5 we analysed the way logged in users navigate the website. For each user we determined the sequence of downloaded pictures and connected two consecutive downloads by a link. These are in fact downloads of the actual pictures rather than just a different index page of a picture. The reason for this selection is that we think that downloading the actual shows that a user is actually interested in the picture itself and not just browsing the lower resolution overview pictures. For each year we superimpose the graphs made by every user, links get counted multiple times if more than one user links the same picture. The result of this analysis can be seen in figure 1. In first approximation we can use the IP addresses of anonymous users to distinguish between and perform the same analysis. The result is shown in figure 2. From 2002 to 2005 the number of nodes in this graph grows from 8069 to 84963. The average number of the node in degree is about 100.

The degree distributions are indicative of scale-free networks, in fact a power law tail with slope 3 as obtained in many models of networks growing with some preferential attachment [6]. We should also investigate other measures to determine the exact nature of the graph. It is certainly surprising to see how little the degree distribution changes with respect to the year.

Currently users are shown the previous and next two pictures as thumbnails in their view. There is also a link to the homepage and overview pages of other

[3] Examples of methods used for automatic text analysis [3] include Latent Semantic Indexing [4].

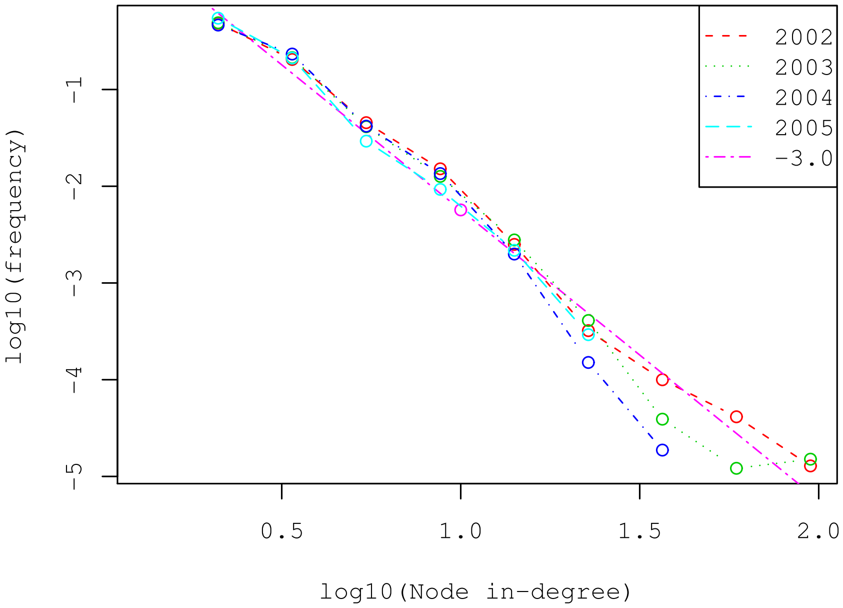

Fig. 1. Distributions of the in-degree of the logged in user graph. The straight line with a gradient of −3.0 appears to be a good fit.

picture in the gallery. Last not least users can change the way pictures from a gallery are filtered by the use of keywords. One of the avenues to be explored by our future research is to find out how this view of the network presented to the user affects the graph their viewing patterns create.

For the performance of the website it is interesting to be able to predict what the most likely picture to be downloaded next is as this could be prefetched. This could cut the delay caused by disk reads on the server side. Another important measure would be to use the data to page-rank [7] the pictures in `gallery.future-i.com`. This could be used to suggest pictures and also for ranking in the search facility.

3 How popular is a picture?

We have seen that the degree distribution of the viewing graph is a power law. As this is closely to the number of downloads per picture we now look at the rank-frequency plot of the download popularity of pictures for the same periods. The results are in 3. As expected the graph shows a power law with a cutoff.

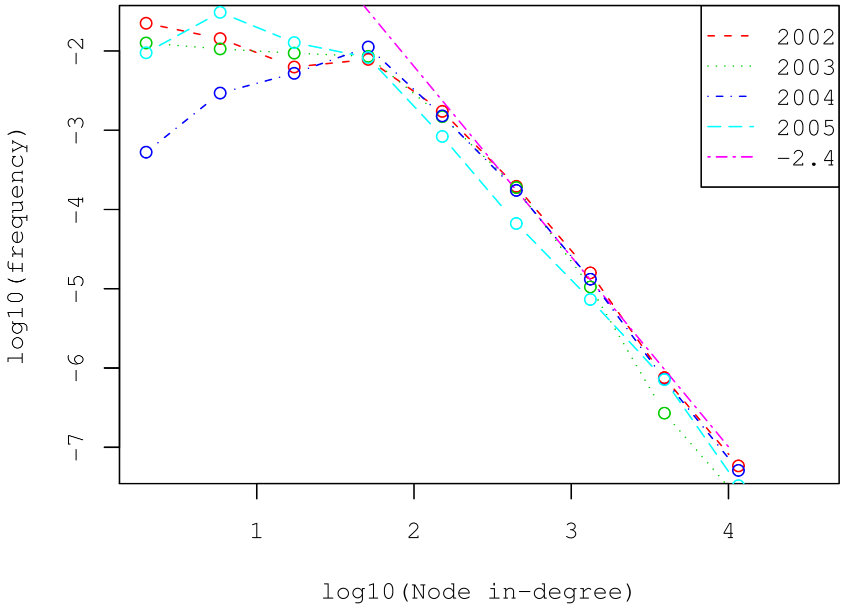

Fig. 2. Distribution of the in-degree of the anonymous users graph. The straight line with a slope of −2.4 is a fairly good fit for the tail of the data.

In fact the graphs shows that the range of the power law increases over time, which could be related to a finite size effect as the number of potential pictures to download increases.

For the most popular picture (`http://gallery.future-i.com/celebs/pic:smallville-kreuk-red-t-montage/`) to download in 2003 we looked at its weekly download rate for the entire observation period. This is displayed in 4. The graph resembles that seen for infection rates on networks [8, 9] with a sharp rise of the "infection rate" which then dies down and carries on as background noise. Similar models have been developed for the spread of rumours [5]. We might be able to infer information about the social network of Gallery users by modelling this data. Similarly we have to look at more pictures and see whether this is indeed a typical behaviour.

4 Conclusion and future work

We have shown that the graph created by successive downloads of picture by individual users has a power law degree distribution. This is indicative of a scale

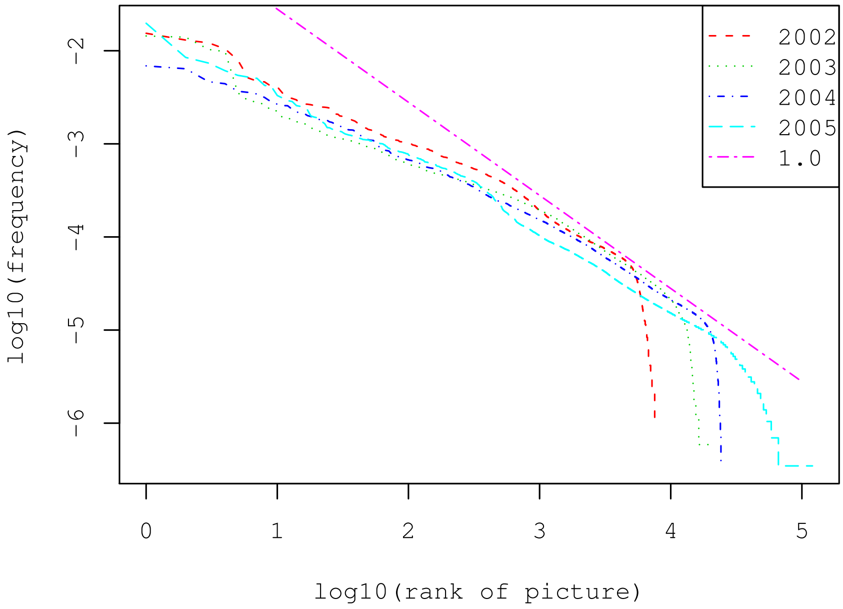

Fig. 3. Rank frequency graphs for the downloaded pictures for four consecutive years. The power law has got a cut-off at about $10^{3.8}$ in 2002 which increases to approximately $10^{4.5}$ by 2005.

free network. Similarly, the rank-frequency plot of picture download shows a Zipf power law. This was not too surprising as the degree distribution of the download graph shows a power law.

There is also evidence that the download behaviour for popular pictures on the `gallery.future-i.com` website shows a similar behaviour as the spread of some infectious diseases and perhaps more appropriately rumours. On the one hand this may be used to infer information of the social structure of the website's user base. On the other hand this might be valuable information when one attempts to model the peak performance of the system under a heavy load..

References

1. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. Nature **435** (2005) 207
2. Bentley, R.A., Lipo, C.P., Herzog, H.A., Hahn, M.W.: Regular rates of popular culture change reflect random copying. Evolution and Human Behavior **In Press, Corrected Proof** (2007) –

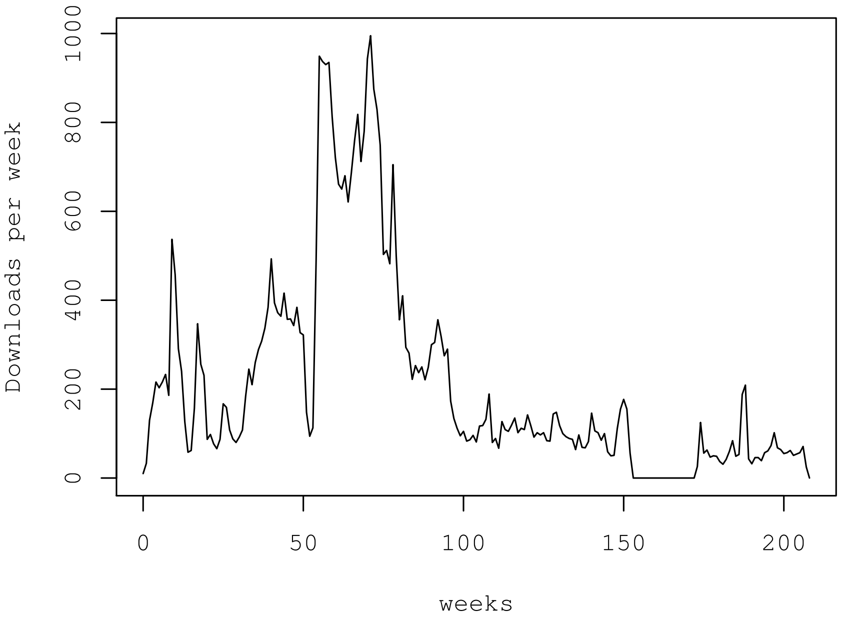

Fig. 4. Anonymous requests per week of the most popular picture in 2003 over almost the entire lifespan of the website.

3. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270 (1994)
4. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41**(6) (1990) 391–407
5. Moreno, Y., Nekovee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Physical Review E **69** (2004) 066130
6. Dorogovtsev, S., Mendes, J.: Evolution of Networks. Oxford University Press (2003)
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
8. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**(14) (Apr 2001) 3200–3203
9. Newman, M.E.J.: The spread of epidemic disease on networks. Physical Review E **66** (2002) 016128