

THE POWER OF ASSUMING NORMALITY

Daphne R. Raban, Eyal Rabin

University of Haifa

Abstract

This study focuses on the power law distributions found on the web and proposes a method to perform statistical inference on data from such distributions. Beyond describing the state of a community, the power law nature of social interactions can be used to explain some of the variance associated with social behavior. Inference based on data on interval or ratio scales rests on the assumption that the data is normally distributed. To obtain normal distributions the power law data is logarithmically transformed and subsequently used in a regression model. Data retrieved from the Google Answers service is used as an example. The regression model suggests that participation in the Google Answers information market is catalyzed both by social and by economic incentives with the most influential incentive being tip, a form of socially-driven economic incentive. This type of analytical approach is seldom found in the internet research literature and is hereby recommended as a very useful analysis tool.

Keywords: Power Law Distributions, Social Networks, Statistical Analysis, Incentives for Participation.

1 INTRODUCTION

Power law distributions describe naturally-occurring events such as earthquakes as well as man-made phenomena such as distribution of book titles sold (Newman, 2005). This article focuses on the power law distributions found on the web and proposes a method to perform statistical inference on data from such distributions.

The power law nature of web interactions has been used mainly to explain network topology (Faloutsos, Faloutsos, & Faloutsos, 1999), to describe various web information sharing environments and to show its wide applicability or universality. Some examples include file sharing (Adamic, Lukose, Puniyani, & Huberman, 2001), web site links (Barabasi & Albert, 1999), electronic markets (Adamic & Huberman, 2000), and discussion groups (Ravid & Rafaeli, 2004). Further analytics are usually employed for social network analysis in order to establish the centrality or success of actors in a social network. The same data can be used for the prediction of social behavior but this is usually overlooked.

This paper proposes that, beyond describing the state of a community, the power law nature of social interactions can be used to **explain some of the variance associated with social behavior**. Data retrieved from the Google Answers service is used as an example to explain this claim.

2 THE NATURE OF THE WEB AS A NETWORK

The web is often described as a social space, a network of networks facilitating a wide variety of information sharing activities between people. Information is shared via hyper-linking, tagging, collaborative compositions, conversations, file transfers, recommendation mechanisms and so on. As a computerized environment which enables tracing and documentation of activity the web is an ideal source for the collection of research data. Scientists have long identified this wealth and have produced many interesting observations about the nature of interactions on the web.

A central empirical observation is that the power law distribution characterizes most, if not all, network-based interactions in large groups. A power law distribution is a scale-free, asymmetrical, asymptotic distribution created by preferential attachment (Barabasi & Albert, 1999). Scale-free is a unique attribute of power law distributions (Newman, 2005) and it means that the same network contains nodes or people whose activity differs by three or more orders of magnitude and the network's capacity to grow is virtually infinite. The distribution of wealth is an example of the scale-free attribute: some people have billions of dollars while other may have only thousands. In this example, the difference is six orders of magnitude. Preferential attachment is a process whereby new entrants will prefer to link or attach to "winners" resulting in a small number of nodes with a large number of links and a large number of nodes with a small number of links (Barabasi & Albert, 1999).

In other words, the structure of the web is not random but it is composed of relatively few highly popular sites (also known as hubs) and a large majority of sites with low popularity, as determined by linking or site visits. Other, more popular, names for the power law distribution are Pareto's Principle, Zipf's Law, and, more recently "the long tail" (Adamic, 2000; Anderson, 2006).

3 QUANTITATIVE ANALYSIS OF WEB ACTIVITIES

Quantitative analysis usually begins with calculation of descriptive statistics and the drawing of histograms showing the distribution of the variables. For the present analysis it is important to note that these initial descriptive analytical procedures are done for the natural, untransformed, data collected.

Quantitative statistical inference is usually divided to parametric and nonparametric tests. Parametric tests are generally used for data that are measured using ratio or interval scales. Nonparametric tests are used when data are measured using ordinal or nominal scales, or, alternatively, with ratio or interval data when the number of observations available is very small. The basic assumption underlying parametric tests is a normal or close to normal distribution of the data. When this assumption holds, a regression model can help explain the relationship between two or more variables, can account for some of the variance in the dependent variable, and can suggest prediction or at least a direction for the relationship.

The activity in online social spaces constitutes field data which has not been manipulated. Rigorous causality cannot be inferred in such unobtrusive data, however, a regression model may shed light on non-trivial phenomena occurring naturally in these social spaces. Regression analysis requires a normal distribution of the data, while web data tend to follow a power law distribution. Normalizing positively skewed, power law, distributions is achieved

by logarithmic transformation. The logarithms of the values are normally distributed. Having fulfilled the assumption of normality a standard regression model is readily produced.

Logarithmic transformation pulls outlying data into the normal distribution. Each single unit on a logarithmic scale translates into a ten-fold change on the original scale of the data. Since the transformation is not linear interpretation of the regression model is made easier if all variables are transformed using the same procedure. However, this is not compulsory as long as the assumption of normal distribution is kept for all variables, natural or transformed. The example provided below uses five variables, four power-law-based and transformed, and the fifth variable, with a normal distribution. It should be emphasized that descriptive statistics are calculated for the original, untransformed data.

4 A CASE IN POINT

Google Answers (GA) was a service offered by Google where search experts, known as Researchers (GARs), provided answers to search questions for a fee offered by the asker. Beside paid answers the system enabled a free exchange of ideas in the form of comments. Another system characteristic was a customer's prerogative to give a monetary tip and/or rate the GAR after receipt of the answer. A more detailed description of the system can be found in Rafaeli et al (2007, forthcoming) or in Edelman (2004) as well as online in the site itself (<http://answers.google.com>). The GA service was discontinued as of December 2006.

GA is an interesting example of a combination an online market and a social space and is unique because the object of trade was information. As will be described in the Method section, distribution diagrams of variables retrieved from GA were mostly classical power law distributions.

Logarithmic transformation and regression analysis was applied to the study of incentives for participation in the Google Answers (GA) information market, or, in other words, to answer the question: what motivates a GAR to provide answers? The unit of analysis was the GAR, the dependent variable was the number of answers provided by the GAR.

5 METHOD

The data from four years (June 2002-May 2006) of GA activity were retrieved and parsed into a database. The original data spanned a longer time period but we excluded two months of activity from the beginning and from the end of the data to avoid incomplete observations. Our database contained 129,745 questions and 52,006 answers given by 523 GARs. The data was summarized on the GAR level since our research interest was in the GAR motivations.

Four independent variables were used to predict one dependent variable. The independent variables were economic and social incentives including: price, tip, rating and number of comments given before an answer was submitted. The dependent variable was the number of answers provided by a GAR.

Price and tip and straightforward numerical values of payment and gratuity offered for the answers that the GARs provided. Price is set when the question is sent by the asker while tip, a voluntary payment, is given after the answer if posted.

Comments and rating require some explaining. Since this study looks at incentives for providing answers by GARs, only comments given before answers were counted. Correlation of the number of such comments to the number of answers is obvious (one cannot have comments before answers when no answer is given); therefore, a proportion of comments-before-answer to answer count was calculated. The abbreviation for this variable is CBPA ("comments before-answer per answer").

The distribution of rating grades (on a five point scale) deviated significantly from normality (skewness=-2.59 indicating a long tail to the left; kurtosis=8.21 indicating a peaked distribution) but was not a power law distribution. Normalization of the rating variable was done by assigning values of zero and one to non-rated and rated answers respectively. The proportion of ratings per answer per GAR was then calculated. This proportion was normally distributed and used for the regression analysis (see Table 2).

Three independent variables (price, CBPA, tip) and the one dependent (number of answers per GAR) variable followed power law distributions. Logarithmic transformation was applied to these variables and the transformed data was used for subsequent regression analysis.

Briefly, the method consisted of:

1. Harvesting the data from the web and deleting incomplete observations.
2. Calculating descriptive statistics.
3. Charting the distributions of the variables selected for analysis and assessing the transformations needed.
4. Transformation of the variables.
5. Checking the distributions of the transformed variables.
6. Performing inferential analysis.

6 RESULTS

Descriptive statistics for the GA site activity appear in Table 1.

Statistic	Current Study
Period of Study	06/2002-5/2006
Duration	48 months
Number of questions asked	129,745
Number of answers provided	52,006
Number of answers with comments	28,034
Number of rated answers	32,854
Number of tipped answers	10,959
Number of experts	523
Average dollar value of question	\$20.90
Average dollar value of answer	\$22.51
Average answer rating (on a 5 point scale)	4.63
Average answer tip value among answers that got tip	\$9.09

Table 1. Descriptive statistics of the GA site activity.

Figure 1 depicts the distribution of number of answers given by GARs. About 17% of the experts provided close to 77% of the answers. The best fit for the data is a power law distribution represented by the following equation:

$$y = 14.51x^{-0.45} \quad (R^2 = 0.64)$$

Figure 2 shows the distribution of the same variable after logarithmic transformation.

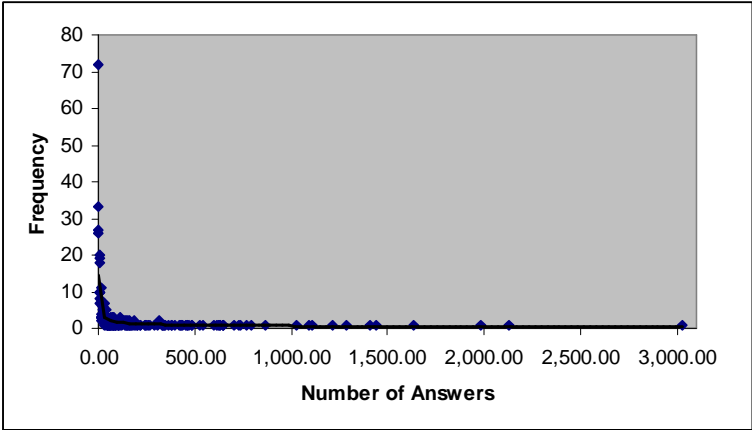


Figure 1. Frequency histogram of the mean number of answers given by GARs.

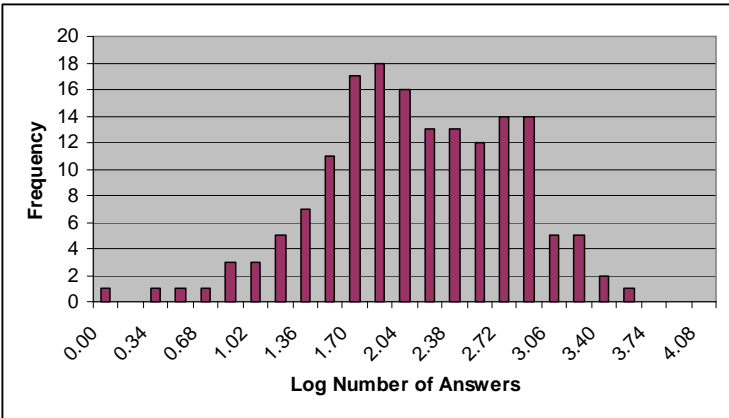


Figure 2. Frequency histogram of the logarithms of mean number of answers given by GARs..

The distributions of price, tip and comments given before the answers followed a similar pattern to the distributions shown in Figures 1 and 2 and are not shown here. Table 2 presents

a summary of the skewness and kurtosis values of the distributions before and after transformation. Skewness is a measure of asymmetry of a distribution. Skewness is zero in a normal distribution. Kurtosis is a measure of how peaked (positive kurtosis value) or flat (negative kurtosis value) a distribution is. In a normal distribution kurtosis is zero.

	Mean	S.D.	Median	Min	Max	Skewness	Kurtosis
Number of Answers	99.43	70.79	13	1	3029	5.60	41.45
Log Number of Answers	1.19	0.84	1.11	0	3.48	0.38	-0.62
Price	17.31	16.78	12.21	2	120	2.70	9.31
Log Price	1.09	0.35	1.08	0.30	2.08	0.08	0.07
CBPA	1.76	1.41	1.78	0	11	1.79	6.84
LogCBPA	0.38	0.22	0.44	0	1.08	-0.38	-0.06
Tip	4.94	0.02	3	0	50	2.73	10.86
LogTip	0.51	0.48	0.60	0	1.71	0.22	-1.32
Rated Answers	62.82	180.49	7	0	2096	6.22	51.11
Proportion of Rated Answers	0.58	0.25	0.61	0	1	-.60	0.49

Table 3. A summary of descriptive statistics, skewness and kurtosis values for the independent and dependent variables before and after logarithmic transformation.

Pearson's coefficients were calculated and regression analysis was performed following logarithmic transformation. Table 3 provides Pearson's coefficients between the four independent variables and the dependent variable. All correlations were statistically significant.

	Log # of Answers	Log Price	Log Tip	Log CBPA	Rating proportion
Log # of Answers	1.00	.42**	.68**	.37**	.14**
Log Price		1.00	.46**	.13**	.08*
Log Tip			1.00	.31**	.22**
Log CBPA				1.00	.07*
Rating proportion					1.00

Table 3. Pearson's correlation values. * $p < .05$. ** $p < .001$.

The regression model was statistically significant ($F_{4,518} = 135.25$, $p < .001$) with an adjusted R square = .507. The predictor variables are shown below:

Variable	Beta	<i>p</i>
Log Price	.14	<.001
Log Tip	.57	<.001
Log CBPA	.18	<.001
Rating proportion	.00	ns

The Beta values indicate the unique contribution of each predictor variable as an incentive for GARS to provide answers.

7 DISCUSSION

Research on the power law distributions characterizing social networks focuses mainly on the structure of the social networks and the implications of the structure. This study uses social network data to perform statistical analysis in order to explain some of the variance associated with behavior patterns.

Specifically, this research looked at the GA information market and investigated the incentives that underlie GARS' participation in the market.

Similar to earlier studies of web-based social networks, most of the variables used here followed power law distributions. Logarithmic transformation produced normal distributions enabling further parametric statistical analysis.

Results show that the main incentive for participation in the GA information market is tip, followed by CBPA and price. Rating was not a significant predictor in the regression model. The emergence of tip as a leading incentive is particularly interesting since tips represent a socially-driven monetary reward (Regner, 2005).

In addition, tipping is done when the value of the answer provided is known to the asker. Thus tipping emerges as an efficient solution to a well-known problem regarding the value of information, namely, the "inspection paradox". This paradox occurs because the limited transparency of information necessitates inspection in order to formulate a value judgment, however, a user cannot inspect information and, in good faith, return it claiming to know nothing (Van Alstyne, 1999). Due to the inspection paradox customers of information are likely to err in their a-priori value judgments for information. The GA pricing system enables offering a moderate price when the question is asked and then providing supplemental payment in the form of a tip when the answer is known to the customer. The relation between pricing and tipping in GA warrants further research.

The overall incentive structure is composed of a mixture of pure monetary incentives (price), socially-driven monetary incentives (tip), and pure social incentives (CBPA). The strongest incentive was the socially-driven monetary incentive, tip. It is suggested that tipping be investigated in additional web markets to evaluate its relation to other incentives and develop theory for and more efficient markets of experience goods (Nelson, 1970; Shapiro & Varian, 1999).

In summary, statistical inference of online social networks must begin with the examination of the original data distributions. Assuming normality without actually checking for it, as is sometimes the case with ratio or interval scales, may undermine statistical analysis. Social interaction patterns often display a power law distribution which requires that descriptive statistics be calculated from the original data, but inference should be based on logarithmic transformation of the data. This type of analytical approach is seldom found in the internet research literature. We recommend it as a very useful analysis tool.

8 REFERENCES

- Adamic, L. A. (2000). Power-law distribution of the world wide web. *Science*, 287(5461), 2115-2115.
- Adamic, L. A., & Huberman, B. A. (2000). The nature of markets in the world wide web. *Quarterly Journal of Electronic Commerce*, 1(1), 5-12.
- Adamic, L. A., Lukose, R. M., Puniyani, A. R., & Huberman, B. A. (2001). Search in power-law networks. *Physical Review E*, 64(4), 46135.
- Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. New York: Hyperion.
- Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Edelman, B. (2004). *Earnings and ratings at google answers*. Retrieved March, 2005, from <http://cyber.law.harvard.edu/people/edelman/pubs/GoogleAnswers-011404.pdf>
- Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. *Computer Communications Review*, 29, 251-262.
- Nelson, P. (1970). Information and consumer behavior. *The Journal of Political Economy* 78(2), 311-329.
- Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), 323-351.
- Rafaeli, S., Raban, D. R., & Ravid, G. (2007, forthcoming). How social motivation enhances economic activity and incentives in the google answers knowledge sharing market. *International Journal of Knowledge and Learning*, 2(4), TBA.
- Ravid, G., & Rafaeli, S. (2004). A-synchronous discussion groups as small world and scale free networks. *First Monday*, 9(9), http://firstmonday.org/issues/issue9_9/ravid/index.html.
- Regner, T. (2005). Why Voluntary Contributions? Google Answers! *CMPO Working Paper Series*(Working Paper No. 05/115).
- Shapiro, C., & Varian, H. R. (1999). *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press.
- Van Alstyne, M. W. (1999, December 12-15, 1999). A proposal for valuing information and instrumental goods. *International Conference on Information Systems*.