Statistical Inference from Power Law Distributed Web-Based Social Interactions

Introduction

Power law distributions describe naturally-occurring events such as the magnitude of earthquakes, the diameter of moon craters or the complexity of the nervous system (Albert Lazslo Barabasi & Albert, 1999; Newman, 2005). Many man-made phenomena also tend to exhibit power law distributions, for example, the distribution of book titles sold, the sizes of cities or intensity of wars (Newman, 2005). Social networks, such as co-authorship networks of scientific collaborations or Hollywood actor networks and web-based social spaces, all follow power law distributions (A. L. Barabasi et al., 2002; Ravid & Rafaeli, 2004). This article focuses on the power law distributions found in web-based social spaces and proposes a method to perform statistical inference on data from such distributions in order to explain behavior and social phenomena.

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 (Albert Lazslo Barabasi & Albert, 1999), electronic markets (Adamic & Huberman, 2000), electronic mail messages (Ebel, Mielsch, & Bornholdt, 2002), discussion groups (Ravid & Rafaeli, 2004) and

response times in computer-mediated communication (Kalman, Ravid, Raban, & Rafaeli, 2006). Further analytics are commonly employed for social network analysis in order to establish the centrality or success of actors in a social network or to discuss weak ties and strong ties (Granovetter, 1983). This kind of analysis usually shows that social standing is not random and is unevenly distributed providing a description of a state but not necessarily an explanation of the actions of actors in the network. Social network analysis has so far concerned itself mainly with the network and group level more than with the individual actors (Wasserman & Faust, 1994). Fairly little social research explored the data derived from networks for the description or explanation or prediction of social behavior. As will be explained below, the studies published to date have used partial data sets while the present article offers a method to analyze complete data sets.

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 and provide a predictive regression model. In order to do that one must analyze the full range of the distribution rather than a sample. Data retrieved from the Microsoft's Netscan project is used as an example to illustrate this claim. We begin by presenting literature on the power law distributions on the web and continue to explain why data from these distributions requires transformation before analysis. Then we describe the analytical procedure and the results using the Netscan data example and suggest future applications of the analysis.

Power Law Distributions on the Web

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 writing, conversations, file transfers, recommendation mechanisms and so on. These various applications raise questions about what motivates people to exhibit certain patterns of participation. 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 interesting observations about the nature of interactions on the web ranging from information sharing behavior (Rafaeli & Raban, 2005) through topics like trust (M. D. Smith, Bailey, & Brynjolfsson, 2000) and cognition (Rafaeli, Raban, & Kalman, 2005) to social aspects of electronic commerce (Rafaeli & Noy, 2005) and value of information (Raban, 2007), to name a few.

A central empirical observation about the structure of social networks, including online networks, is that the power law distribution characterizes network-based interactions in large groups. A power law distribution is a scale-free, asymmetrical, asymptotic distribution (Albert Lazslo Barabasi & Albert, 1999). Scale-free is a unique attribute of power law distributions (Newman, 2005) which means that the same network contains nodes or people whose attribute of activity differs by orders of magnitude, but they are all described by the same distribution. For example, people can send email messages that are 10 kilobytes or 1 megabyte in size, and they will still belong to the same

distribution. Scale-free distributions have several important attributes including: the distribution contains a small number of very large nodes and a huge number of small nodes, the network's capacity to grow is virtually infinite, it is quite resistant to displacement of nodes (as most of them are small) but vulnerable to changes in major hubs or nodes.

The mathematical formula representing the power law distribution is given by:

$$p(X) = aX^{-\alpha}$$

Say that X stands for the number of links pointing to a certain web site or to the activity in online forums. The probability of measuring X varies inversely as a power of X. Another way to express the inverse relationship is to assign a negative sign to the exponent. The formula indicates that the probability of large events is very small and the probability of small events is high. Power law distributions of social networks are typically attributed to a process called preferential attachment or positive feedback 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 (Albert Lazslo Barabasi & Albert, 1999; Shapiro & Varian, 1999).

The exponent α is often in the range of 2.1 to 4, but is not limited to that range (Albert Lazslo Barabasi & Albert, 1999; Newman, 2005). Power law distributions are detected by plotting data on log-log axes, which is the same as taking the log (base 10) from both variable sets, X and p(X). A straight line on

the plot indicates a power law distribution with the slope of the line equaling the exponent, α:

$$LOG\mathit{p}(X) = -\alpha LOGX + LOGa$$

In other words, there is a linear relationship between the log-transformed values of X and p(X). Focusing on the power law structure of the web hyperlinks as an example, this structure is not random but it is composed of relatively few highly popular sites or nodes (also known as hubs) and a large majority of nodes or sites with low popularity, as determined by linking or site visits. Additional, more specific or more popular, names for the power law distributions are Pareto's Principle, Zipf's Law, and, more recently "the long tail" (Adamic, 2000; Anderson, 2006; Newman, 2005).

Quantitative Analysis of Web Activities

Parametric quantitative analysis generally deals with normally distributed, randomly sampled data. Correlation and regression analysis are based on an assumption of co-linearity between variables analyzed which usually holds for such data.

As mentioned before, the power law distribution is anything but normal, and variables exhibiting this distribution often do no display a linear relationship among them even in cases where such a relationship is theoretically expected based on the content of the variables. For example, in a discussion forum or a Usenet group one would expect a close to linear relationship between the number of messages and the number of participants in the group. It seems trivial

or natural to expect a simple relationship where a rise in the number of participants in a Usenet group will result in a correspondingly higher number of total messages. Our data indicate differently.

In the following we propose how to analyze data originating from social interactions on the web that have power law distributions. A linear relationship expected between variables such as the number of messages and number of participants does not exist in the original data. Descriptive statistics show us that the distribution is power law. Since we know that in power law equations we have a linear relationship between the log transformed (base 10) values of X and of p(X) performing this transformation leads to co-linearity between these variables. The log transformation enables correlation and linear regression analyses. Moreover, it enables using the *full* set of observations, in contrast to previous research that examined partial sets of data (Fiore, Tiernan, & Smith, 2002; Peddibhotla & Subramani, 2007; M. D. Smith et al., 2000; Soroka & Rafaeli, 2006). Reducing the set of observations based on the magnitude of one variable leads to range restriction which may lower the validity of the findings. In one of the articles just cited descriptive statistics were offered on the partial set and in the other cases linear regression was performed on a partial set. Such analytical procedures may lead to an inaccurate representation of data as we show in a hypothetical example that follows.

Hypothetical Power Law Data Example

To illustrate why it is inaccurate to analyze partial data when it is power law distributed we use a hypothetical example. In this example we take a set of

X values: all the natural numbers between 1 and 100. We find their corresponding p(X) values by using the power law equation and the value of

α=-2.5 as the exponent. Figure 1 shows the data (up to X=40) set and the equation. It is quite clear visually that co-linearity between X and p(X) will be low using the entire data set, but could be high when using a sub-set of data, especially for high values of X, which may give an inaccurate representation of the relationship. This may give the wrong impression that the assumption of co-linearity required for regression is filled, but it is not. In this hypothetical example the correlation coefficient for the entire data set is about -0.22 while the correlation coefficient for a sub-set of X values ranging between 21-40 is about 0.95. This indicates clearly that selecting a limited range of data points as a sub-set for analysis, as was done in some of the earlier cited studies, is not representative of the entire population and should not be done.

Take in Figure 1

Figure 1: Power law distribution of hypothetical data per the equation shown.

Being natural and unobtrusive the full data reflect the actual behavior of individuals or the aggregated behaviors of participants in online communities; therefore, performing analysis on the full data is expected to find a variety of applications in social and behavioral research performed in web settings including a multitude of the so-called Web 2.0 applications. The following section proposes how to analyze full sets of power-law distributed data.

Statistical Inference and the Power Law Distribution

As a rule, the basic assumption underlying parametric statistical tests is that the data is normally (or close to normally) distributed. This assumption can be somewhat relaxed with large sample sizes. Social networks on the web are usually very large consisting of dozens, hundreds, or even thousands of members. However, their distributions are often power law, which are very different from normal or close-to-normal distributions. The mean and the median in power law distributions are very distant from each other (highly skewed), and the distribution has no maximum value (high kurtosis value). Performing statistical inference on natural data which displays a power law distribution will lead to mis-estimation of the effect since only a subgroup of the data may conform to the basic assumptions. Adding to the confusion is the statistical significance which will usually appear despite the problems mentioned because of the large number of observations.

One method of parametric inference is regression analysis which 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 assumptions of linear regression include: a. there is a linear relationship between the variables. b. the random errors of each variable are normally distributed. c. the random errors of each variable are independent. d. homoscedasticity (constant variance) of the errors of the dependent variable for each value of the independent variable.

As explained earlier, the logarithm of power law data yields a straight line so performing a logarithmic transformation of the data offers an easy solution to the problem of analyzing power law distributed data. Each single unit on a logarithmic scale translates into a ten-fold change on the original scale of the data. Logarithmic transformation stretches the upper end of the distribution and compresses the tail so its overall effect is to reduce the kurtosis and skewness values. Since the transformation is not linear interpretation of the regression model is made easier if all variables are transformed using the same procedure. Nevertheless, this is not compulsory. In the case of regression analysis the important point is to keep the assumption of co-linearity for all variables, natural or transformed.

In summary, the activity in online social spaces constitutes field data which has not been manipulated providing a unique opportunity to understand social processes. 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 is based mainly on the assumption of co-linearity of the data, while web data tend to follow a non-linear, power law distribution, and the relationship between variables is not linear. The following section provides an example based on Microsoft's Netscan project where most activity of participants fits a power law distribution. The aim in this example is to study motivations for behavior using the variables available on the Netscan web site. Logarithmic transformation is performed prior to regression

analysis. It should be emphasized that descriptive statistics are calculated for the original, untransformed data.

A Case In Point: Netscan

Netscan is a vast database constructed by Microsoft Research, Community Technology Group (http://netscan.research.microsoft.com). The database documents the activity in Usenet newsgroups and is described by Microsoft as follows:

"The Netscan System provides detailed reports on the activity of Usenet newsgroups, the authors who participate in them, and the conversation threads that emerge from their activity. Using the Netscan tool users can get reports about any newsgroup for any day, week, month, quarter, or year, since September 1999."

Netscan has been used to research the value of authors (Fiore et al., 2002), to visualize conversations (M. A. Smith & Fiore, 2001; Turner, Smith, Fisher, & Welser, 2005), and to differentiate between frequent and infrequent users (Brush, Wang, Turner, & Smith, 2005). These intriguing research projects, based on data mining techniques coupled with subjective measures, reveal rich interaction patterns and roles of different community members. The present report uses Netscan data to demonstrate the proposal for analyzing power-law distributed data using multiple regression.

Netscan contains a large amount of data in three levels: aggregated newsgroup data, author participation metrics and thread development. For the

present research aggregated newsgroup data was used. Specifically, the data selected covered the activity in all newsgroups that had the word 'computer' in their name during the month of January 2007. The data was representative to the extent that other similar sets of data were assumed to follow power law distributions. This was checked randomly and was found to be true. The present research does not attempt to obtain statistical sampling such that will enable to generalization of the findings. The aim here is to generalize the method of analysis, not the findings.

The main variables available at the aggregated newsgroup level are defined in Table I. All these variables exhibited power law distributions.

Take in Table I

Table I: Definition of variables at the aggregated newsgroup level

It seems trivial that there should be a linear relationship between the number of posts and the other variables in Table I; however, as shown in the Results section, plotting the data (Figure 3 in the Results section) reveals that a linear relationship is very weak and virtually does not exist.

Using the Netscan data as an example for using data which follows power law distributions to explain behavior (beyond describing network structure and characteristics) we present the following research question: What are the antecedents of posting messages in a computer-related Usenet newsgroup?

The purpose is to elucidate which of the independent variables are the best predictors for the number of posts in Usenet groups. The Method section details the variables selected and the method of analysis.

Method

Using the Netscan search option data was retrieved for the month of January 2007 aggregated at the newsgroup level for all groups that had the word 'computer' in their names. A total of 550 newsgroups were thus retrieved with data saved in a spreadsheet and a statistical package for further analysis.

Five independent variables were used to predict one dependent variable. The dependent variable was "Posts" (see Table I) and the independent variables were Posters, Returnees, Replies, Repliers, Average Line Count.

The method of analysis consisted of:

1. Harvesting the data from the web.

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 co-linearity of the transformed variables as well as the other assumptions for regression analysis.

6. Performing regression analysis.

Results

Descriptive statistics for selected data originating from Netscan appear in Table II. Power law distributions are highly skewed as implied by the high values of standard deviations in Table II.

Take in Table II

Table II: Means and ranges of variables in Netscan 'computer' Newsgroups for January 2007

Figure 2 depicts the distribution of the number of posts per computer-related newsgroup (the dependent variable). The best fit for the data is a power law distribution represented by the following equation:

$$y = 21{,}931x^{-1.43} \ (r = 0.956)$$

While the exponent here is somewhat smaller than the earlier cited range of 2.1-4, the data distribution is definitely power law as evidenced by the line equation and the high correlation coefficient, and also as determined by a curve estimation procedure.

Take in Figure 2

Figure 2: Power law distribution of the Posts per Group in Netscan computer-related newsgroups.

The distributions of all the independent variables, posters, returnees, replies, repliers and average line count, followed a similar pattern to the distribution shown in Figure 3 and are not shown here. Table III summarizes the power law equations and correlation coefficients obtained by the curve fitting procedure for each of the variables in this study.

Take in Table III

Table III: Power law equations for all variables

None of the independent variables had a linear relationship with the dependent variable in their natural form. Figure 3 depicts the relationship between Posts and Posters in Netscan computer-related newsgroups as an example of a non-linear relationship.

Take in Figure 3

Figure 3: Relationship between the dependent variable, Posts, and one of the independent variables, Posters (r=0.144)

The linear correlation between the two variables in Figure 3 is very poor as evidenced by the correlation coefficient (r=0.144). Linear regression is

meaningless since the assumption of co-linearity does not hold. Similar low correlations emerged for all five independent variables and the dependent variable (Posts): Posters (r=0.144), Returnees (r=0.130), Replies (r=0.055), Repliers (r=0.052), Average Line Count (r=0.029).

As cited earlier, when power law data is plotted on a log-log chart, a straight line is obtained. Accordingly, Figure 4 shows the log-log plot for the same variables which appeared in Figure 3, namely the relationship between the logarithm (base 10) of the dependent variable, log Posts, and the logarithm (base 10) of one of the independent variables, log Posters in Netscan computer-related newsgroups (r=0.881).

Take in Figure 4

Figure 4: Relationship between log Posts and log Posters (r=0.881)

The linear correlation between the two variables in Figure 4 is good as evidenced by the correlation coefficient (r=0.881). Logarithmic transformation allows proceeding with Pearson's correlations and a linear regression model, which are both based on the assumption of a linear relationship between the variables.

Pearson's correlation coefficients were calculated and regression analysis was performed following logarithmic transformation. Table IV provides Pearson's correlation coefficients between the five independent variables and the

dependent variable. All correlations with the dependent variable (log Posts) were statistically significant.

Take in Table IV

Table IV: Pearson's correlation values for the logarithms of the variables

In order to avoid multi-co-linearity we performed the regression using the stepwise method. The model was statistically significant *($F_{3,549}$=913.62, p<.001)* with an adjusted R^2=.833. The predictor variables and their beta values in the regression model for predicting the average number of posts per computer-related Usenet group in Netscan are shown in Table V below:

Take in Table V

Table V: Linear regression (stepwise) outcomes

The Beta values indicate the unique contribution of each predictor variable in the regression model for predicting the average number of posts per computer-related Usenet newsgroup in Netscan during the month of January 2007.

Without the logarithmic transformation the adjusted R^2=.019 (*$F_{1,549}$=11.53, p<.001)*; Posters were a statistically significant predictor with a Beta value of .144 (*p*<.001) and all the other possible/hypothesized predictors were excluded from

the regression model. The logarithmic transformations improved the goodness of fit of the equation and help to explain more of the variance of the dependent variable.

Finally, the assumptions for performing a linear regression analysis require an inspection of the errors. Histograms of the residuals of all the variables were found to exhibit a normal distribution. A residuals analysis revealed that only 11 out of 550 observations in the data set were outliers (2%) above two standard deviations, which is less than the 5% allowed.

Discussion

Research on the power law distributions characterizing networked social spaces focuses mainly on the structure of the social networks and the implications of the structure. The present study used data on the *activity* in a social network to perform statistical analysis in order to explain some of the variance associated with behavior patterns. Previous studies that analyzed motivations for behaviors in web-based data with power law distributions did not use logarithmic transformations. Those studies analyzed only a subset of the data or sufficed with descriptive statistics (Fiore et al., 2002; Peddibhotla & Subramani, 2007; M. D. Smith et al., 2000; Soroka & Rafaeli, 2006). Using those strategies compromises the wealth of data available. The claim in this paper is that despite the nature of power law distributions characterizing social activity, statistical inference can be performed under certain conditions, and this offers many opportunities for analyzing social influences in the online environment both

on the web and in organizational networks such as intra-nets and knowledge management systems.

Using a hypothetical data set we showed why using the full range of data is important for correct representation of the relationship between an independent and a dependent variable. Then, by using specific data harvested from Microsoft's Netscan project we demonstrated the feasibility of the analysis suggested. The same analysis can later be applied to additional datasets to obtain generalizability of the specific findings; however, this was not the purpose of the present research. The purpose here was to demonstrate the methodology using Netscan data as an example. Specifically, the goal of the analysis was to determine the antecedents of posting in computer-related Usenet newsgroups in order to elucidate which of the independent variables were the best predictors.

Similar to earlier studies of web-based social networks cited above, all of the variables used here followed power law distributions and there was no linear relationship between the independent and dependent variables. Logarithmic transformation (base 10) produced co-linearity enabling further parametric statistical analysis, correlation analysis and multiple regression.

Results show that the main influences on the number of posts in computer-related Usenet groups during the time period of the current study (January 2007) were the number of posters and the number of returnees (returning authors) with a slight contribution by the average line count. The number of participants (posters, returnees) were the main predictors of activity (posts) rather than the extent of the activity as measured by replies. This is an

interesting observation that warrants further research and theory development. We suggest that a possible theoretical approach to the research of Usenet groups' activity may be to analyze and compare social capital (presence of people) with cultural capital (exchange of information and knowledge).

In a way, the main finding, that the number of posters is the best predictor for the number of posts, sounds trivial. One wonders if social research is needed to make that prediction. This is exactly the point of this article, namely, to show that this trivial relationship does not exist for the raw data when they follow a power law distribution. After logarithmic transformation this natural relationship unfolds, as expected. While this paper focuses on methodology and benefits from producing a seemingly trivial finding, it shows the rigor of the methodology which opens up rich opportunities for social research on data originating from networks. This data may come from social networks such as LinkedIn or Facebook, it may be retrieved from social tagging systems like Del.ici.ous, from information sharing sites like Youtube, and a host of other user-generated content applications. Similar techniques of analysis may be applied for studying intra-organizational processes since social technologies (wikis, blogs, tagging) are becoming popular within organizations.

In the current study the regression model yields that 83.3% of the variance in the dependent variable, posts, is explained by three of the five independent variables listed in Tables I and II. The importance of this finding is its predictive power in the regression of the logarithmic values. To determine the original

values, one needs to convert the predicted logarithmic values by raising them to the power of ten (anti-log).

As mentioned earlier, the purpose of this paper is to propose a method of analysis and not to thoroughly research behavior patterns in Usenet groups. The method proposed here enables the pursuit of such a study. In Netscan researchers can use this analytical framework to study seasonal effects on participation, individual effects, topical effects, interactivity, silence and more aspects that may influence participation. Large data sets may be divided into subsets in order to investigate whether motivations of highly populated nodes are similar to motivations driving the less populated nodes (Raban, 2008). In fact, any so-called "Web 2.0" web site, meaning sites where intensive social and information exchange interactions take place, is an excellent candidate for analysis by the method presented here. One must find a way to harvest data from social sites, a task easily achieved by capable programmers, and then examine and analyze the data with standard software. In addition, corporations wishing to study the performance of participatory sites on their intra-nets or which is often a part of knowledge management systems may apply the proposed framework.

In summary, statistical inference of online social networks must begin with the examination of the original data distributions. Descriptive statistics are extracted from the original data which is subsequently logarithmically transformed to enable correlation and regression analysis. Assuming co-linearity without actually checking for it, as is sometimes the case with large samples

based on 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 information systems or Internet research literature. We recommend it as a very useful analysis tool.

Acknowledgements

Supported by a grant from the Israel Foundations Trustees (2006-2008).

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.

Barabasi, A.L., Jeong, H., Nٰda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, 311(3-4), 590-614.

Brush, A.J.B., Wang, X., Turner, T.C., & Smith, M.A. (2005). Assessing differential usage of usenet social accounting meta-data. Proceedings of the SIGCHI conference on Human factors in computing systems, 889-898.

Ebel, H., Mielsch, L.I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. Physical Review E, 66(3), 35103.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. Computer Communications Review, 29, 251-262.

Fiore, A.T., Tiernan, S.L., & Smith, M.A. (2002). Observed behavior and perceived value of authors in usenet newsgroups: Bridging the gap. CHI Letters, 4(1), 323-330.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. Sociological Theory, 1, 201-233.

Kalman, Y., Ravid, G., Raban, D.R., & Rafaeli, S. (2006). Pauses and response latencies: a chronemic analysis of asynchronous CMC. Journal of Computer-Mediated Communication, 12(1), http://jcmc.indiana.edu/vol12/issue11/kalman.html.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323-351.

Peddibhotla, N.B., & Subramani, M.R. (2007). Contributing to public document repositories: A critical mass theory perspective. Organization Studies, 28(3), 327-346.

Raban, D.R. (2007). User-centered evaluation of information: A research challenge. Internet Research, 17(3), 306-322.

Raban, D.R. (2008). The Incentive Structure in an Online Information Market. Journal of the American Society for Information Science and Technology, 59(14), 2284-2295.

Rafaeli, S., & Noy, A. (2005). Social Presence: Influence on Bidders in Internet Auctions. EM - Electronic Markets, 15(2), 158-176.

Rafaeli, S., & Raban, D.R. (2005). Information sharing online: a research challenge. International Journal of Knowledge and Learning, 1(1/2), 62-79.

Rafaeli, S., Raban, D.R., & Kalman, Y. (2005). Social Cognition Online. In Y. Amichai-Hamburger (Ed.), The social net: The social psychology of the internet. Oxford, England: Oxford University Press.

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.

Shapiro, C., & Varian, H.R. (1999). Information Rules: A Strategic Guide to the Network Economy. Boston: Harvard Business School Press.

Smith, M.A., & Fiore, A.T. (2001). In Visualization components for persistent conversations (pp. 136-143). Paper presented at the SIGCHI conference on Human factors in computing systems, Seattle, Washington. ACM.

Smith, M.D., Bailey, J., & Brynjolfsson, E. (2000). Understanding digital markets: Review and assessment. In E. Brynjolfsson & B. Kahin (Eds.), Understanding the Digital Economy: Data, Tools, and Research (pp. 99–136). Cambridge, Mass.: MIT Press.

Soroka, V., & Rafaeli, S. (2006). In Invisible participants: How cultural capital relates to lurking behavior. Paper presented at the WWW 2006, Edinburgh, Scotland.

Turner, T., Smith, M.A., Fisher, D., & Welser, H.T. (2005). Picturing Usenet: Mapping computer-mediated collective action. Journal of Computer-Mediated Communication, 10(4).

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications: Cambridge University Press.

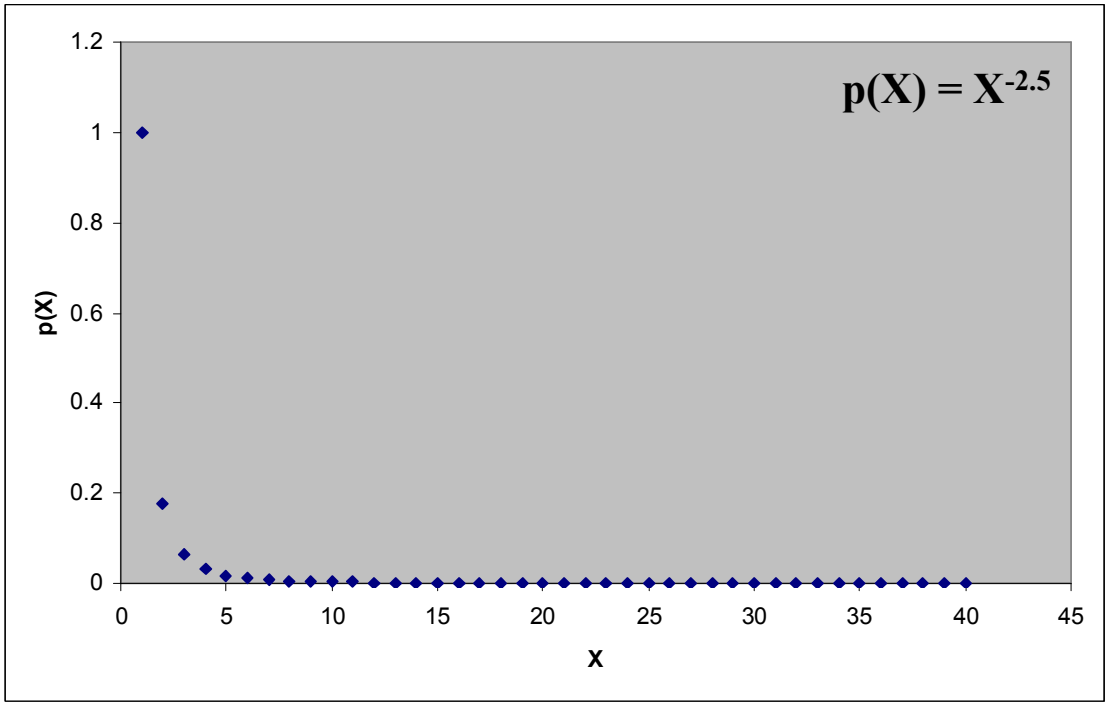

Figure 1: Power law distribution of hypothetical data per the equation shown.

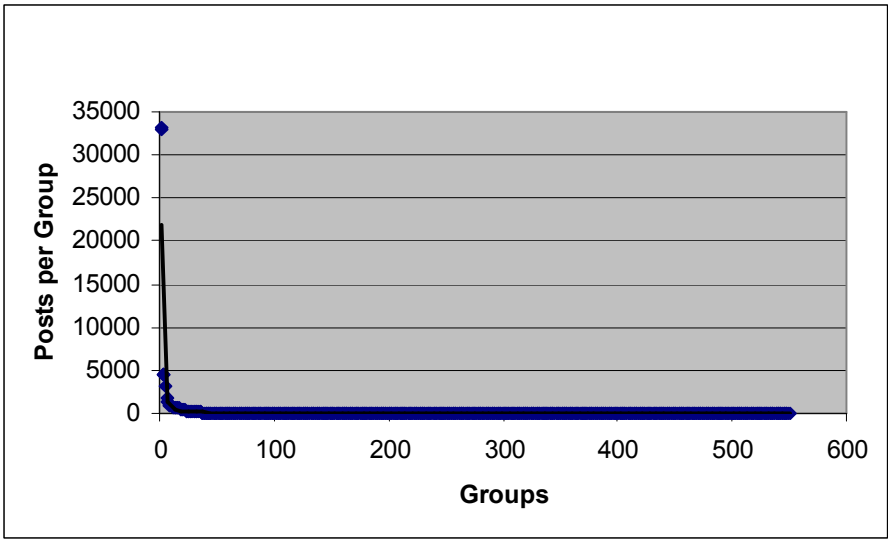

Figure 2: Power law distribution of the Posts per Group in Netscan computer-related newsgroups.

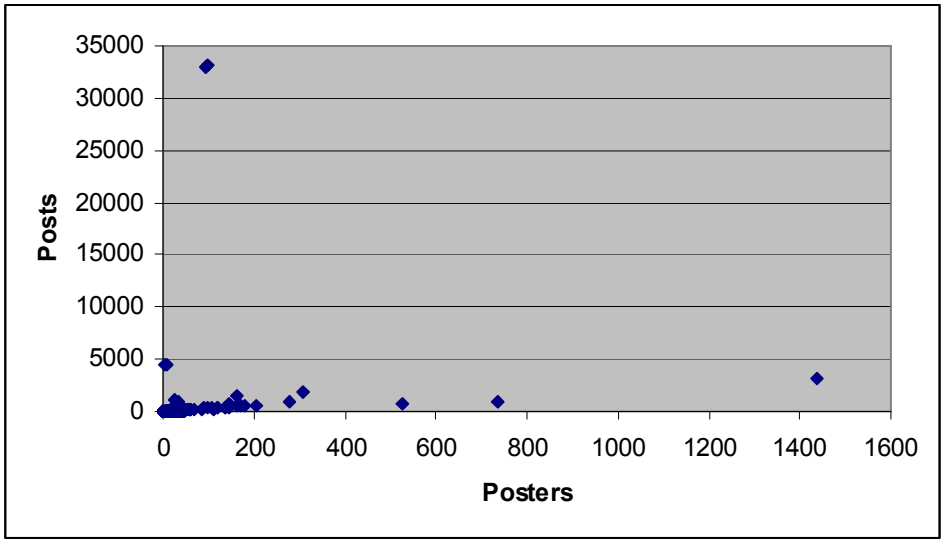

Figure 3: Relationship between the dependent variable, Posts, and one of the independent variables, Posters (r=0.144)

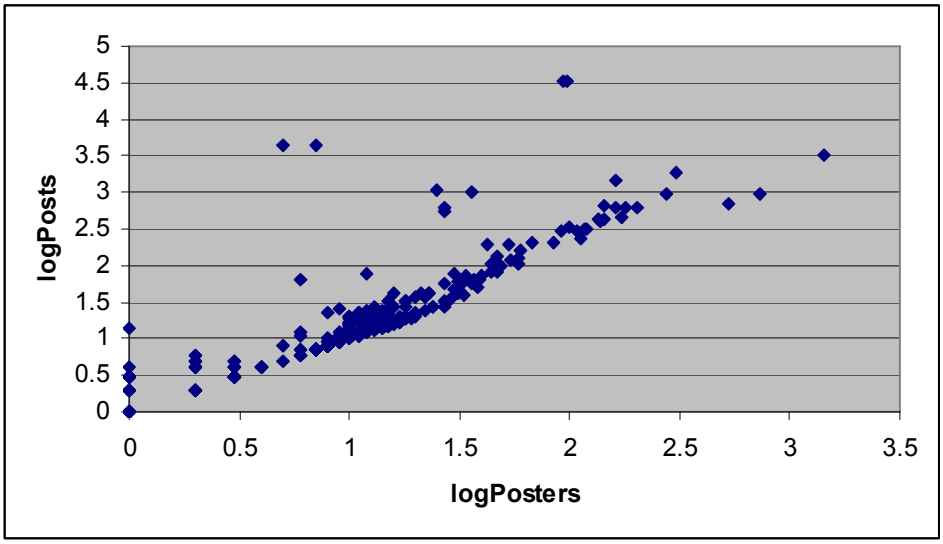

Figure 4: Relationship between log Posts and log Posters (r=0.881)

Variable	Meaning
Posts	Number of messages in the newsgroup
Posters	Number of authors in the newsgroup
Returnees	Number of posters who contributed in the previous time period
Replies	Messages that were sent in reply to other messages
Repliers	Posters who replied at least once
Average Line Count	Average line count per group

Table I: Definition of variables at the aggregated newsgroup level

Variable	Mean	Std Dev	Minimum	Maximum
Number of posts per group	184.44	2018.49	1	33,217
Number of posters per group	20.11	77.06	1	1,436
Number of returnees per group	2.84	11.46	0	114
Number of replies per group	16.82	108.99	0	1,636
Number of repliers per group	4.39	19.69	0	244
Average number of lines	45.55	153.82	2	2,783

Table II: Means and ranges of variables in Netscan 'computer' Newsgroups for January 2007

Variable	Equation	r
Posts	$y = 21{,}931x^{-1.426}$	0.956
Posters	$y = 2172.4x^{-1.053}$	0.888
Returnees	$y = 745.57x^{-1.326}$	0.975
Replies	$y = 24224x^{-2.060}$	0.958
Repliers	$y = 2419.8x^{-1.582}$	0.944
AvgLineCt	$y = 795.28x^{-0.629}$	0.908

Table III: Power law equations for all variables

	Log Posts	Log Posters	Log Returnees	Log Replies	Log Repliers	LogAvg LineCt
LogPosts	1	.881**	.765**	.635**	.634**	.276**
LogPosters		1	.668**	.577**	.588**	.296**
LogReturnees			1	.822**	.827**	.118**
LogReplies				1	.990**	-.031
LogRepliers					1	-.034
LogAvgLineCt						1

Note: **p<.001

Table IV: Pearson's correlation values for the logarithms of the variables

Variable	B	Std.Error	Beta	t	p	Cumulative R^2
Log Posters	0.82	0.03	.651	26.52	<.001	.776
Log Returnees	0.52	0.04	.325	13.75	<.001	.881
LogAvgLineCt	0.10	0.04	.045	2.45	<.01	.883

Table V: Linear regression (stepwise) outcomes