Journal of Data Science **11**(2013), 559-573

High-Quality Winners Take More: Modeling Non-Scale-Free Bulletin Forums with Content Variations

Frederick K. H. Phoa* and Wei-Chung Liu
Academia Sinica

Abstract: Modeling the Internet has been an active research in the past ten years. From the "rich get richer" behavior to the "winners don't take all" property, the models depend on the explicit attributes described in the network. This paper discusses the modeling of non-scale-free network subsets like bulletin forums. A new evolution mechanism, driven by some implicit attributes "hidden" in the network, leads to a slightly increase in the page sizes of front rank forum. Due to the complication of quantifying these implicit attributes, two potential models are suggested. The first model introduces a content ratio and it is patched to the lognormal model, while the second model truncates the data into groups according to their regional specialties and data within groups are fitted by power-law models. A Taiwan-based bulletin forum is used for illustration and data are fitted via four models. Statistical Diagnostics show that two suggested models perform better than the traditional models in data fitting and predictions. In particular, the second model performs better than the first model in general.

Key words: Bulletin forums, category-specific degree distribution, non-scale-free network, power-law distribution, preferential attachment.

1. Introduction

Modeling the Internet has been an active research in the past ten years. Most of the models are for scale-free networks that follow the traditional preferential attachment mechanism, and the properties of these networks follow power-law distribution. Most networks are not scale-free in the real world. It leads to an introduction of a baseline generated by an additional uniform attachment from the traditional preferential attachment. A category-specific degree distribution is suggested for the properties of these networks. From the "rich get richer" behavior to the "winners don't take all" properties, these models depend on the explicit attributes described in the network, and they simply assume that any attributes

*Corresponding author.

that are not described by the network edges do not have any significant effects to the evolution of the network. In reality, it is not the common case. A bulletin forum is one of these examples.

Bulletin forums, or simply called forums for the rest of this paper, are online discussion site where people can hold conversations in the form of posted messages. They are originated from a computer system running software called a Bulletin Board System, or BBS, which allows users to connect and log in to the system using a terminal program. The first BBS-like system, which was called Community Memory, was started in 1973 in Berkeley, California (Crosby, 1995). The forty-year technological evolution has changed the BBS from a small-scale dial-up neighborhood communication in 1970s, to a popular larger-scale Telnet-connected system within schools and associations in 1980s, and eventually to a web-based modern forum in 1990s.

Nowadays, forums are considered as web applications managing user-generated content. A sense of virtual community often develops around forums that have regular users. A hierarchical structure of a forum is common for finite content organization, so a forum can also be viewed as a graph or a network (Barabási *et al.*, 1999). However, unlike the Internet that has close to infinite number of users, user-registered systems are found in most forums, and the number of forum users is thus limited to a finite size.

In this paper, we aim at studying the formation and evolution of a general forum, which we view it as a subset of the whole Internet. It is organized as follows. We first discuss two most popular network evolution mechanisms in Section 2. We point out some features that these two mechanisms are unable to describe and we introduce a new mechanism that are plausible drivers of the evolution of the size of the forum page. In Section 3, we introduce a Taiwan-based BBS called PTT as an illustrative example of a forum. We also discuss briefly the truncation idea. In Section 4, we build models on the distribution of the PTT board size ranking. Two potential models are suggested to improve the fitting of this board size ranking data from the traditional two models. Section 5 further discusses the newly introduced mechanism that is applied to two suggested models. Other potential models are suggested at the end.

2. Two Network Evolution Mechanisms for Forum Page Sizes

The size of a forum page, which is generally considered as an important property of a forum network, is defined as the number of users that have actions, e.g., post a topic or paragraph, in the forum page. One of the most documented aggregate network properties is the so-called “scale free” property, whose degree distribution asymptotically follows a power law. This property has been commonly found in computer science, see Broader *et al.* (2000), Crovella and

Bestavros (1997), Faloutsos *et al.* (1999) and etc. Formally, a network is said to have a power-law degree distribution when for degree k, the probability distribution k follows a power-law, i.e.,

$$P(k) \propto k^{-\gamma},$$

where γ is a parameter of the power-law (PL) distribution.

There are many mechanisms that are plausible drivers of the evolution of a general network. In particular, there are several major mechanisms that lead to the distribution of the network of forums. First, we consider preferential attachment, the most common explanation for the emergence of PL degree distributions (Barabási and Albert, 1999, de Solla Price, 1976). Second, we consider category-specific degree (CSD) distribution, which is a generalization of mixing preferential attachment with uniform attachment (Pennock *et al.*, 2002). As we show that these two mechanisms are not enough to characterize the distribution of the size of a forum page, we consider a new mechanism that related to the content of a forum page. Notice that the content covered in a forum page is not shown or based on any network structure, but the users usually enter a forum page with some general and interesting contents.

2.1 Preferential Attachment and Power-Law (PL) Model

Barabási and Albert attribute PL scaling to a "rich get richer" mechanism called preferential attachment (Barabási and Albert, 1999). This mechanism suggests that in a forum network, a user is more likely to be entered to a well-known forum page that has last for a long time. It provides an explanation for the emergence of PL degree distribution. A mathematically rigorous treatment of the preferential attachment model can be found in Bollobás *et al.* (2001).

The preferential attachment captures how the current structure of a network influences the creation of new edges. However, due to the hierarchical structure of a forum, the users generally do not link a forum page from outside, compared to the major link from its previous page. Moreover, if a discussion forum has been established for some significant times, the increase of the number of active users will slow down and eventually hits a limit. This limit can be the number of users who are interested in the specific topics discussed in this forum, or just simply the limit due to the capacity of the server behind the forum. There will only be very few new edge formations when a forum reaches a limit state, so the preferential attachment mechanism has less affects in this situation.

2.2 Category-Specific Degree (CSD) Distribution

Pennock and his co-authors realized the difference between a massive social network and its subsets (Pennock *et al.*, 2002). When a PL scaling is interpreted, the "winners take all" phenomenon suggests that only a few popular pages benefit from a greater exposure to the public and the majority of sites have a difficult time competing for user attentions. However, at small connectivity, Pennock and his co-authors observed a growing divergence between the distribution of links and the fit from PL distribution. The discrepancy is larger for outbound links than for inbound links.

In order to fix this discrepancy, a generative network growth model is suggested that every vertex has at least some baseline probability of gaining an edge instead. Therefore, the endpoints of edges are chosen according to a linear combination between the preferential attachment and uniform attachment at an adjusted probability. The resulting distribution, called CSD distribution, consists of a roughly lognormal body and a power law tail. Since only a subset of the whole Internet is considered, the CSD model provides a better fit than the PL distribution governed singly by preferential attachment.

2.3 Effects on Hidden Attributes

Both the preferential attachment and its mixture with a uniform attachment are edge formation mechanisms based on the structure of the network. They ignore the influence from some attributes of a forum page (e.g., the content covered or the topic's interest level) that are not based on the network structure but they may affect the forum network evolution. However, most of the forums are characterized by some specific topics and the users who joined the forums usually have special interests in these specific topics discussed in the forums. These hidden attributes lead to some deviations from the lognormal body of the CSD distribution. Although some statistical models are available to capture such mechanism, like the exponential random graph models (Robins *et al.*, 2007, Handcock *et al.*, 2008) and the SIENA model (Snijders, 2001, 2005), there has been little work on how these "hidden" attributes influence the evolution of large network.

In the forum network, the forum page that covers more contents and/or more interesting topics should be more likely to attract users to read and perform actions. Furthermore, if a forum has a specific theme, then a forum page will be more attractive if it provides information closely related to the theme. These forum pages are generally called "hot pages".

3. Description and Analysis Method of PTT: A Taiwan-based Bulletin Forum

PTT Bulletin Board System, or simply called PTT, is a terminal-based BBS based in Taiwan. The main site was found in 1995 and it became the largest online forum in Taiwan since 2000. Nowadays, this BBS, using the telnet protocol, is arguably the largest BBS in the world with more than 1.5 million registered users. During peak hours, there are over 150k users online.

Our data consists of all board names and their sizes in a monthly basis for four months in 2010. If we treat the whole PTT as a forum, then a board in a PTT is equivalent to a forum page, and the size of a board is defined as the number of PTT users that has actions in the board. Notice that the PTT boards can be further classified into public and private boards. Our data include public boards only, because the private boards are owned by individual group of users and the use of these private boards without permissions violates the privacy of these users. Therefore, boards are referred as the public boards from now on.

PTT is possibly the largest forum that can be found in the Internet, so we expect some features of the Internet exist in PTT. However, PTT still possesses an important forum property that the number of users is limited. Another problem is that it is very difficult to quantify the content coverage and topic interest for every PTT board, because the PTT users come from a wide spectrum of different purposes and backgrounds. Instead, we assume that the PTT boards can be classified into several groups according to their board size rank, and the amount of contents covered and the topics being interested by users are the same. This leads to a truncation of the whole PTT boards into several groups and different models or model parameters are used to fit the board size ranks in different group of pages.

4. Data Analysis

According to our data, PTT does not have significant growth in size as the number of boards maintains at about 2700 for all four months. Therefore, we may build a static model for the PTT board size ranking. In this section, we treat the average of February to April 2010 data as our training data and the May 2010 data as our testing data for comparison. The distribution of the ranking of PTT board size is shown in Figure 1. In specific, Figure 1 (left) is the distribution of the training data (February 2010 - April 2010) and Figure 1 (right) is the distribution of the testing data (May 2010). We perform the data fitting via the PL and the CSD models for reference, and then our suggested models are proposed and compared with the first two traditional models.

4.1 Fitting Training Data with the Power Law (PL) Model

The ranking of PTT board size possesses similar characteristics to some general frequency-rank data. The median size of these boards is 45 but the one

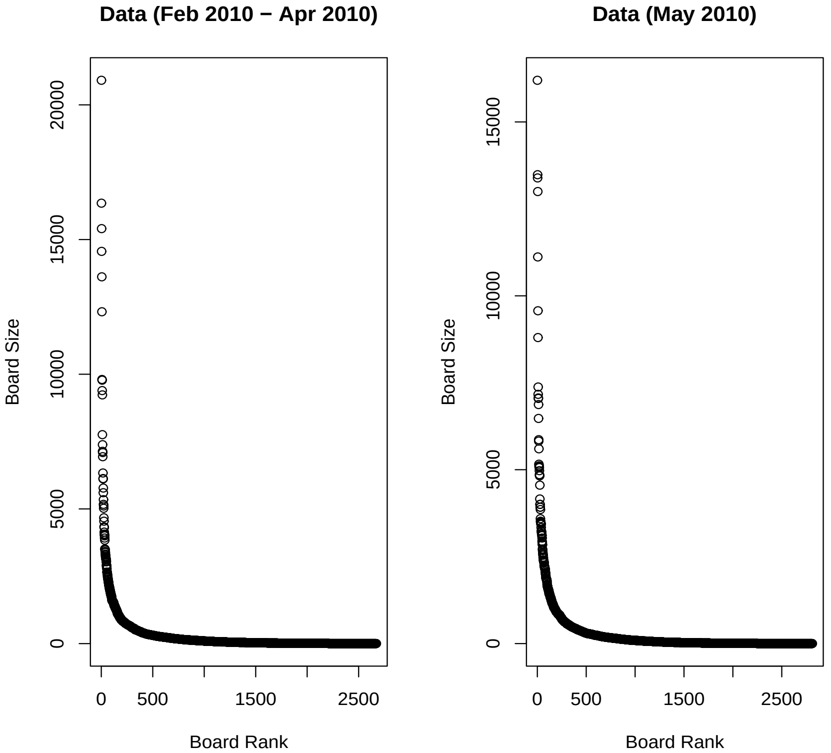

Figure 1: The distribution of PTT Board Size Ranking Monthly Data. It consists of PTT board names and its corresponding board size. (Left) This training data is collected from February 2010 to April 2010. The data point in this figure is the average of each board in three months. (Right) This testing data is collected in May 2010

with largest size reaches 20910. In fact, 85% of the sizes of all PTT boards come from the top 18% largest boards, so the 80-20 principle approximately holds in our data. A PL model, or a Zipf's distribution, is standard to describe the distribution of ranked data that 80-20 principle holds (Barabási and Albert, 1999). A regression on the log-log scale of the data suggests that

$$\log \hat{S} = 16.79 - 1.88 \log k,$$

or equivalently

$$\hat{S} = 19607913 k^{-1.88},$$

where $\hat{S}$ and k are the estimated size and the rank of the PTT board. Figure 2 (top left) shows the fitting using the above equation in original and log-log scales. It is obvious that the fitting of the PL model cannot capture the characteristics of our data.

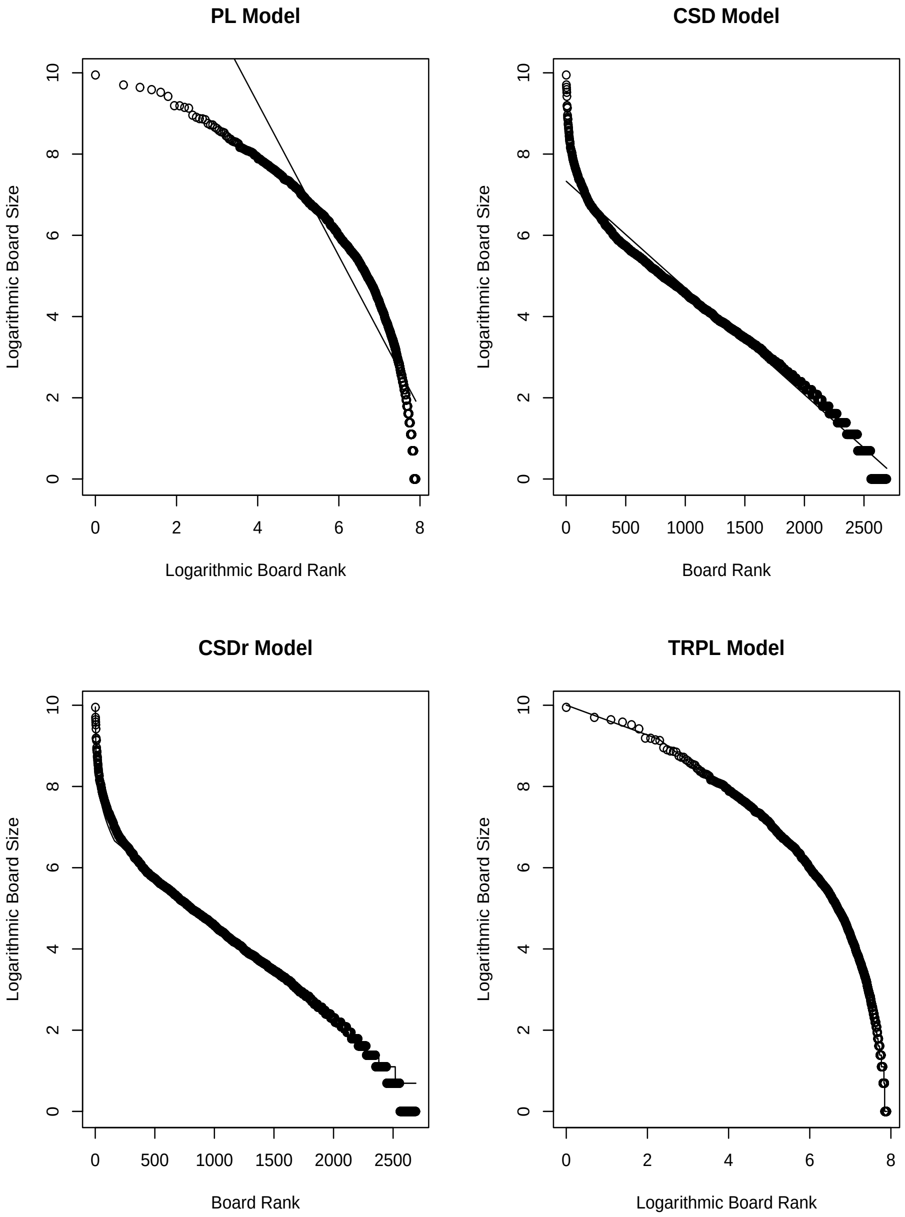

Figure 2: PTT Board Size Ranking Data fitted by four different models. (Top left) Data is fitted by Power Law (PL) model, and the figure is plotted in log-log scale. (Top right) Data is fitted by Category-specific Degree (CSD) model, and the figure is plotted in logarithmic scale. (Bottom left) Data is fitted by Category-specific Degree with content ratio (CSDr) model, and the figure is plotted in logarithmic scale. (Bottom right) Data is fitted by Truncated Regression Power Law (TRPL) model, and the figure is fitted in log-log scale

4.2 Fitting Training Data with Category-specific Degree (CSD) Model

Due to the discrepancy between the data and power-law model, a CSD model is suggested in Pennock *et al.* (2002), which considers the edge formation is under the mixture mechanism of preferential attachment and uniform attachment. A regression on the logarithmic scale of data suggests that

$$\log \hat{S} = 7.33 - 0.0026k,$$

or equivalently

$$\hat{S} = 1529e^{-0.0026k}.$$

Figure 2 (top right) shows the fitting using the above equation in original and logarithmic scales. The CSD model fit our data better than the PL model, but it still underestimates the size of the front rank boards.

4.3 Category-specific Degree with Content Ratio (CSDr) Model

The CSD underestimates the sizes of the front rank boards, mainly because these hot boards are usually the main feature of the whole forum and they contain the hottest and the most general contents that are highly related to the theme of the forum. A general CSD model cannot capture the exceptionally high number of users having actions in these hot boards, so an adjustment ratio for content is introduced for these hot boards. The first 128 PTT boards are classified as hot in every month by default. However, since the list of hot boards changes every month, where some previously hot boards drop off from the list and some newly hot boards enter into the list, we keep track of the top 6% of boards (162 boards) and treat them as hot boards in general.

An adjustment ratio $\hat{R}$ is needed to be multiplied to the baseline lognormal size estimates $\hat{S}$ in order to describe the hotness of the boards. Therefore, using boards other than the top 6% boards, a regression on the logarithmic scale of data suggests that the estimate of baseline board size is

$$\log \hat{S} = 7.054 - 0.0024k \Leftrightarrow \hat{S} = 1159e^{-0.0024k}.$$

The ratio itself can be treated as a ranked data. Therefore, using the hot boards, a regression on the log-log scale of data suggest that

$$\log k_0\hat{R} = 2.9430 - 0.5709 \log k \Leftrightarrow \hat{R} = \frac{26.50}{k_0}k^{-0.57},$$

for $k = [1, 162]$ and $k_0 = 1.3969$ (the size of the last board in the second group) is the normalization for segment continuity. Therefore, the final CSDr model can be written as

$$\hat{S} = \begin{cases} 21980k^{-0.57}e^{-0.0024k}, & \text{if } k \leq 162, \\ 1159e^{-0.0024k}, & \text{if } k > 162. \end{cases}$$

Figure 2 (bottom left) shows the fitting using the above equation in original and logarithmic scales. The CSDr model provides an uprising trend for the front rank PTT boards, which fits our data better than the original CSD model.

4.4 Truncations of PTT Boards into Groups: A Simple Way to Deal with Content Variations of the Boards

There are boards with various amount of contents in PTT. The content variations mainly comes from the lack of restrictions of board formations among PTT administrators. This leads to the current situation that some boards contain broad range of contents but some have narrow and specific scopes. When a network data like this is analyzed, instead of the preferential attachment, this content variation, which is a hidden attribute in the network, potentially becomes a major mechanism to the PTT board size. Therefore, some statistical tools are needed to minimize the effect of this hidden attribute.

A truncated regression divides the whole data into several segments according to its regional specialties. In our PTT data, we divide our data based on the board size. We assume that two PTT boards have similar content coverages if their board sizes are close. Then an appropriate truncation of our data basically removes the content variations of PTT boards. For our data, we decide to divide it into six different categories, corresponding to six different groups of PTT boards with different content covered, topics being interested and ways of linkages.

1. Top 10 boards. They are the boards with the largest sizes, or in other words, the number of users that visits these boards is extremely large when compared to other boards. In fact, 15.76% of the sizes of all PTT boards come from these 10 boards.

2. Top 6% boards. They are the boards with the largest 6% sizes except the top 10 boards (a total of 152 boards in our data). In fact, 46.30% of the sizes of all PTT boards come from these 152 boards. Together with the top 10 boards, they are frequently considered as hot boards and have additional links from the front page for users to visit.

3. Boards with insignificant sizes. They are the boards that their sizes are less than 1% of the size of the smallest top 6% board. In our data, the smallest

top 6% board has size 1091, so any boards that have size less than 10.91 will belong to this category.

4. Top-tier and bottom-tier boards. The sizes of these boards are in between those of top 6% boards and boards with insignificant sizes. In our data, their size ranges from 11 to 1091. We divide this large subset of boards into two groups. The boards in the front 50% ranks belong to top-tier boards and those in the rest 50% ranks belong to bottom-tier boards.

5. Boards with no responses. They are the boards with size 1, which means no one responds after the first message is posted. There are 135 boards with size 1 in our data.

Note that the choice of "Top 10 boards", "Top 6% boards" and "3% noise threshold" are all arbitrary, but a slightly change on the choices do not severely affect the result in the modeling. For example, the results are similar if "Top 5 boards", "Top 5% boards" and "5% noise threshold" are chosen instead.

4.5 Truncated Regression Power-Law (TRPL) Model

Although the CSDr model seems to capture the feature in the front rank boards, some local discrepancies and the overestimates of the tail both suggest that lognormal body is unable to capture all features of the data, and perhaps the PL model with careful data truncation may achieve a better fit. Therefore, for the first five groups, we follow our assumption that boards in each group have similar content coverages. In addition, it is safe to further assume that the uniform attachment is so small in these five groups that we can simply ignore them. Then our PTT board sizes in the first five groups are fitted with PL distributions with different parameters. For the last category, since it is barely considered as a board due to the lack of interactions with other PTT users, we simply estimate the number of boards by the average number across the months of training data.

By considering the log-log scale of the data in the first category, the size of the top 10 boards can be modeled in the following PL distribution.

$$\hat{S}_1 = e^{10.00}k^{-0.37},$$

for $k = 1, \cdots, 10$. The model has adjusted $R^2 = 0.9236$ and p-value$= 5.98 \times 10^{-9}$. The similar approach is used in modeling board size ranking in the second to the fifth categories. This leads to the description of our data via a truncated regression model as follows.

$$\hat{S} = \begin{cases} \hat{S}_1 = 2209k^{-0.37}, & \text{for } k = [1, 10], \\ \hat{S}_2 = 5168k^{-0.74}, & \text{for } k = [10, 162], \\ \hat{S}_3 = 76529k^{-1.26}, & \text{for } k = [162, 829], \\ \hat{S}_4 = e^{22.26}k^{-2.56}, & \text{for } k = [829, 1494], \\ \hat{S}_5 = e^{43.38}k^{-5.43}, & \text{for } k = [1494, 2256], \\ \hat{S}_6 = 1, & \text{for } k = [2556, 2691]. \end{cases}$$

The adjusted R^2 of the first five models are 0.9236, 0.9960, 0.9923, 0.9960 and 0.9752 respectively, and the p-values are very close to zero in all five models. Figure 2 (bottome right) shows the model fitting of our suggested model in the original and log-log scales.

4.6 Diagnostics on the Simulated Probability Distributions

The probability density of our data can be described using the above truncated regression model in the following way.

$$P_T(k) = \frac{\hat{S}_i(k)}{\sum_{i=1}^{6} \hat{S}_i(k)}.$$

We compare the goodness of fit between the traditional power-law distribution and the probability density suggested by our truncated regression model using the root mean square error (RMSE) and the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). In particular, RMSE aggregates the individual differences of points between the true and estimated density into a single measure of predictive power. Mathematically speaking,

$$RMSE = \sqrt{E((\hat{S} - S)^2)},$$

where $E(\cdot)$ is the expected value of the difference. KL divergence is a non-symmetric measure of the difference between two probability densities and it is

$$D_{KL}(P_{true} \| P_{est}) = \sum_k P_{true}(k) \log \frac{P_{true}(k)}{P_{est}(k)},$$

where $P_{true}(k)$ and $P_{est}(k)$ represents the true and the estimated probability densities.

First, we examine our model via the training data (the average of February to April 2010 data) and the statistics are given in Table 1 (top). The values of RMSE and the KL divergence sorted in the descending order are PL model, CSD model, CSDr model and TRPL model. Clearly the probability densities via TRPL and CSDr models have less difference to the true density than the two traditional models in the training data, and the estimates of TRPL model is closer to the true density than CSDr model.

Table 1: Diagnostics of PL, CSD, CSDr and TRPL models via the RMSE and the Kullback-Leibler (KL) divergence

PTT Training Data Estimations:		
	RMSE	KL Divergence
PL Model	1.08×10^{-2}	2.4629
CSD Model	9.62×10^{-4}	0.2895
CSDr Model	2.49×10^{-4}	0.0206
TRPL Model	7.80×10^{-5}	0.0013
PTT Testing Data Predictions:		
	RMSE	KL Divergence
PL Model	1.13×10^{-2}	2.5807
CSD Model	8.63×10^{-4}	0.2405
CSDr Model	2.83×10^{-4}	0.0021
TRPL Model	1.89×10^{-4}	0.0052

Then, we use four models to predict the testing data, which is the PTT board size ranking of May 2010 and the statistics are given in Table 1 (bottom). The traditional two models are still not as good as our two suggested models in both RMSE and KL divergence. In addition, TRPL model seems to perform better than the CSDr model, only except a slightly higher KL divergence in the testing data prediction. Notice that the current data truncation is arbitrary, and we believe a more careful truncation will lead to a better fitting. In summary, the TRPL model performs the best followed by the CSDr model.

5. Discussion

This paper suggests three mechanisms for the evolution of forum properties. The first mechanism is the well-known preferential attachment, the second mechanism is the mixture of uniform attachment specific for the subset nature of the network, and the third mechanism relates to the evolution of attributes that are not expressed in the network. We believe that in our PTT data and perhaps many non-scale-free bulletin forum data, the third mechanism is the major force instead of the preferential attachment and the uniform attachment, which is different from the conventional wisdom about the evolution of properties in the Internet.

This difference mainly comes from the structure of the forum itself, or more generally speaking, the partial structure inside the Internet. Internet is a huge network such that most of the properties can be considered as in the infinite size, like the number of users, number of clicks in a page, and so on. The partial structure of the Internet, on the other hands, has various restrictions or limitations. For our PTT data, even though it is arguably the largest BBS in the world, the number of users, especially active users, is only a tiny part of the whole Internet population. This specific property leads to a finite limitation about the number of users when we model the data.

Due to the limit number of users, preference of each user becomes relatively important. This leads to the third mechanism, where a forum page that covers general, hot and related-to-the-theme content gains more attentions. Therefore, the first potential modification is done on the content ratio adjustment of the CSD model, and it leads to the CSDr model. In fact, the so-called content ratio is a complicated quantity and it consists of many "hidden" attributes possibly governed by many sociological and psychological mechanisms. Since it is too complicated to analytically derive this quantity, a data-driven approach may approximate the quantity. However, as we see in our PTT data, the fitting is not fully satisfied.

The truncated regression model, on the other hands, tries to eliminate the effects of these "hidden" attributes by dividing all forum pages into several groups. Then an assumption about the no-effects of these "hidden" attributes is introduced. Therefore, by grouping the pages, the baseline uniform attachment and the content effects are eliminated within groups. It is good enough to fit the Zipf distribution, which is originated from the preferential attachment, for the ranked data within groups. However, it is obvious that a more sophisticated or even theory-driven truncation approach is needed in order to optimize the data fitting, and it is still under investigation.

There are other potential modeling techniques that can be considered instead of the CSDr and TRPL models. For example, it is possible to divide the forum pages into different group according to their topics, instead of the relative ranking. The model can be fitted under the assumption that each user has a specific preference or hobby to read. Then for each group, it is possible to fit a Zipf or lognormal distribution in each group and hopefully there exists only one big page and the sizes decay exponentially over ranks within groups. In this sense, a mixture model is resulted when these distribution are combined.

This paper uses RMSE and KL divergence to compare the fitness among four models. Although both criteria are well-known measures, the values of the two measures are sensitive to probability values of distribution. It is not recommended to randomly sample boards from the population (2691) boards to form sampling distribution and then calculate the RMSE and KL divergence, because the sampling process easily drops large boards. Consider the data used in this paper, there are 10 out of 2691 boards that have large board sizes. There are about 15.5% probability of dropping these 10 large boards in a random sample of 500 boards. Instead of sampling, it is recommended to obtain more monthly PTT data and calculate RMSE and KL divergence on these newly obtained data.

Acknowledgements

The study leading to this paper is supported by Academia Sinica, Taiwan, and is a part of research project titled "Modeling the Structure and Dynamic of Large Social networks: Taiwan-based Data Development and Analysis" (Grant Number: AS-100-TP2-C01-3). The authors would like to thank the administrators of PTT for providing valuable data to analyze, and Ms. Pei Wen Tsai for data cleaning process.

References

Crosby, K. (1995). Convivial cybernetic devices. *The Analytical Engine* **3**(1).

Barabási, A. L., Albert, R. and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physical A* **272**, 173-187.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000). Graph structure in the web: experiments and models. In *Proceedings of the 9th International World Wide Web Conference*, 309-320. Elsevier Science, Amsterdam, New York.

Crovella, M. E. and Bestavros, A. (1997). Self-similarity in World Wide Web traffic evidence and possible causes. *IEEE/ACM Transactions on Networking* **5**, 835-846.

Faloutsos, M., Faloutsos, P. and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Proceedings of the ACM SIGCOMM 1999 Conference*, 251-261. ACM Press, New York.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-512.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**, 292-306.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. and Giles, C. L. (2002). Winners don't take all: characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences* **99**, 5207-5211.

Bollobás, B., Riordan, O., Spencer, J. and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structure and Algorithm* **18**, 279-290.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks* **29**, 173-191.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). Statnet: software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software* **24**, 1-11.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. In *Sociological Methodology* (Edited by M. E. Sobel and M. P. Becker), 361-395. Basil Blackwell, Boston.

Snijders, T. A. B. (2005). Models for longitudinal network data. In *Models and Methods in Social Network Analysis* (Edited by P. Carrington, J. Scott and S. Wasserman), 215-247. Cambridge University Press, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

Received November 8, 2012; accepted March 4, 2013.

Frederick K. H. Phoa
Institute of Statistical Science
Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan
fredphoa@stat.sinica.edu.tw

Wei-Chung Liu
Institute of Statistical Science
Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan
wliu1975@stat.sinica.edu.tw