Analyzing and ranking the Spanish speaking MySpace community by their contributions in forums

Andreas Kaltenbrunner
andreas.kaltenbrunner@barcelonamedia.org

Rafael Banchs
rafael.banchs@barcelonamedia.org

Erika Bondia
erika.bondia@barcelonamedia.org

Barcelona Media – Innovation Centre,
Av. Diagonal, 177, 08018 Barcelona, Spain

ABSTRACT

This study analyzes the MySpace forums in Spanish language, which permits to extract otherwise restricted demographic data from the social network and leads to a detailed description of the Spanish speaking community in MySpace. Important differences in age and gender structures are found for users with different countries of origin. The distribution of activity among different threads and users is very heterogeneous and follows heavy tailed distributions. We furthermore propose two variants of the h-index which allow the ranking of users and threads by their importance in the forums.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Human Factors, Languages, Measurement, Algorithms.

Keywords

social network, forum, post, online demography, h-index

1. INTRODUCTION

As the web 2.0 continues its fast expansion and consolidation, social networks and virtual communities are becoming more and more popular. This phenomenon is certainly changing the way humans interact in both, the cyberspace and the real world, and has attracted the attention of research communities from many different areas of knowledge. Indeed, in the last few years, there has been a lot of research work involving some relevant issues of social networks. In some cases, the studies have focused on determining mathematical models for understanding and explaining the structure of social networks [16, 13], as well as the dynamics of people interactions and communications [12, 10] or the relationship between personal behavior and social group affiliation [23], while others concentrate mainly on friendship networks [9, 24, 5].

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Nevertheless, despite this large amount of scientific activity around the social network phenomenon, most studies are centered on English speaking communities; and, as far as we know, no comprehensive study has been carried out yet for a Spanish speaking collective. This motivated us to focus here on a Spanish and Latin American part of a social network.

We selected for this study the Spanish speaking community of MySpace[1] and, more specifically, the study focuses on forum threads and posts created by this collective. MySpace was created in 2003, and has become the social network with the highest number of registered users, with roughly 148 million of actives users [8]. In spring 2007, MySpace launched its Latin version for Hispanic users. And since then it was possible to use MySpace in Spanish. However, in Spain it was not officially presented until June of 2007, although the months before it was already possible to use it in form of a beta version. Nowadays, according to recent studies [25] MySpace is also the leader social network in Spain, with a share of 34% of all users that participate in any social network. The MySpace forums consist of a fixed set of topics (i.e. the forums), where users can create new threads and posts to them relative freely with little restrictions (contrary to the English speaking forum which sometimes have a more restrictive moderation system).

A previous study of MySpace has been already carried out in [24] but in contrast to our work, it was centered on the entire community of MySpace, without any preference of language, and was focused on the friendship relations among users. The study suffered from a large amount of private profiles in the dataset. To avoid this possible distortion of the results we used only data which is public for all users which participate in the forums.

The present work has been conceived with two main objectives in mind. First, we intend to generate a reference study for MySpace's Spanish speaking community, which we expect will provide a baseline for future comparative studies about the evolution of this virtual community, as well as, the structural differences between the English and Spanish speaking communities of MySpace. Second, from a more methodological point of view, we propose two measures based on the h-index [11] (normally used to rank scientists) as metrics for relevance ranking, not only for threads, but also for the users in a virtual community.

The paper is structured as follows. First, section 2, describes the data collection process and some statistical fun-

[1] `http://www.myspace.com`

damentals used in the study. In section 3 we describe the main statistics and distributions of the collected data and section 4 presents a brief description of the two proposed metrics, h-thread and h-user, for relevance ranking of threads and users, respectively. Finally, section 5 presents our main conclusions and ideas for future work.

2. METHODS

2.1 Data retrieval

We retrieved the entire amount of forum posts in Spanish available at MySpace at May 16th, 2008 in form of raw HTML-pages. The total amount of retrieved data was 1.7GB which was transformed into XML for further processing and imported into Matlab where the statistical analysis was performed. The XML-files contain the following data for every thread: thread-id, title of the thread, user-id (of its creator), time-stamp (when the thread was started) and the text of the initial post. For every post in one of these threads we have the following information: thread-id, post-id, user-id (of its author) and time-stamp (when the post was published). Finally, for every user: user-id, user-name, sex, age, city, province and country of origin.

The oldest posts in our dataset where published on December 13, 2007 and the most newest on the very same day we collected the data.

2.2 Statistical fundamentals

Several quantities in our dataset show heavy tailed distributions [22]. We approximate them with two of the exponents of this family of distributions, **power law (PL)** [19] and **log-normal (LN)** distributions [14], which often create competing models to explain the same data [17].

The LN-distribution has the following probability density function (pdf):

$$f_{LN}(t;\mu,\sigma) \quad = \quad \frac{1}{t\sigma\sqrt{2\pi}}\exp\left(\frac{-(\ln(t)-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

For the PL-distribution we use its discrete version, whose pdf is given by

$$f_{PL}(t;\alpha,t_{min}) = \frac{t^{-\alpha}}{\zeta(\alpha,t_{min})} \qquad (2)$$

where

$$\zeta(\alpha,t_{min}) = \sum_{n=0}^{\infty}(n+t_{min})^{-\alpha} \qquad (3)$$

is the generalized or Hurwitz zeta function [1] and t_{min} the lower bound for the powerlaw behavior.

To find the optimal value for t_{min} to fit a dataset we use a method proposed in [6], which is based on finding the t_{min} for which the distance between the cumulative distribution functions of data and a PL-fit (using maximum likelihood estimation) reaches a minimum value. This is in most cases equivalent to maximizing the p-value of a Kolmogorov-Smirnov (KS) test [7], which we use as well to measure the quality of our PL and LN approximations.

The p-value of the KS-test gives us the probability of obtaining a result as different from the approximating distribution as the data. In other words: the greater the p-value, the closer is the fit with the test distribution. The fit is accepted if the p-value is greater than the chosen level of significance α_0 (usually set to 0.05 or 0.01).

3. STATISTICS

In this section we present some statistical quantities of the data collected. After describing the size of the dataset we focus on the distribution of activity (posts) between different users and threads and finally analyze the length of the users contributions as well as the demographic structure of the Spanish speaking collective in MySpace.

3.1 Global statistics

Our dataset contains about 300000 posts produced by about 23340 distinct users. The posts are divided into approximately 25000 threads. For more details see Table 1.

Number of threads	24923
Number of users	23340
Number of forums	22 primary (79 secondary)
Number of posts	298048

Table 1: Principal quantities of the data collected.

In Table 2, we summarize some of the maximum amounts observed in the analysis. The most surprising fact is the existence of a thread with more than 15000 posts, which represents more than 5% of all posts observed in the study and a user with more than 8500 posts.

Largest thread	15499 posts
Most users in the same thread	303 users
User with most posts	8577 posts
User in most threads	1833 threads
User with most threads initiated	725 threads

Table 2: Maximum values of some of the variables studied.

3.2 Heterogeneous behavior

When analyzing the distribution of posts among the different users and threads in our dataset we obtain a very heterogeneous picture of activity in the forums.

For instance, the 300000 posts are far from being equally distributed among the threads. Figure 1 top shows the distribution of the number of posts per thread. It has the typical form of a heavy tailed distribution. Which in this case means that although the majority of threads contains few posts, there is a considerable amount with a very large number of posts, compared to the average or the median. A thread here receives on average 12 posts, but 47.8% of the threads receive just one post (which implies that the median is close to 1). However, about one in 100 threads receives more than 100 posts.

This distribution of the data (black circles in Figure 1) can be approximated well with a log-normal distribution (continuous red line) as can be seen in the top right plot of Figure 1, which shows the cumulative distribution of the number of posts per thread. The good visible fit of the data is confirmed by the high p-value of 0.41 of a KS-test. One can also adjust the tail of the distribution with a power law (blue dash-dotted line) with an exponent of -2.2 and a corresponding p-value of 0.83. However, this fit is only acceptable in the white area of Figure 1, which implies that one has to discard 93.2% of the threads (gray area) when adjusting a power law.

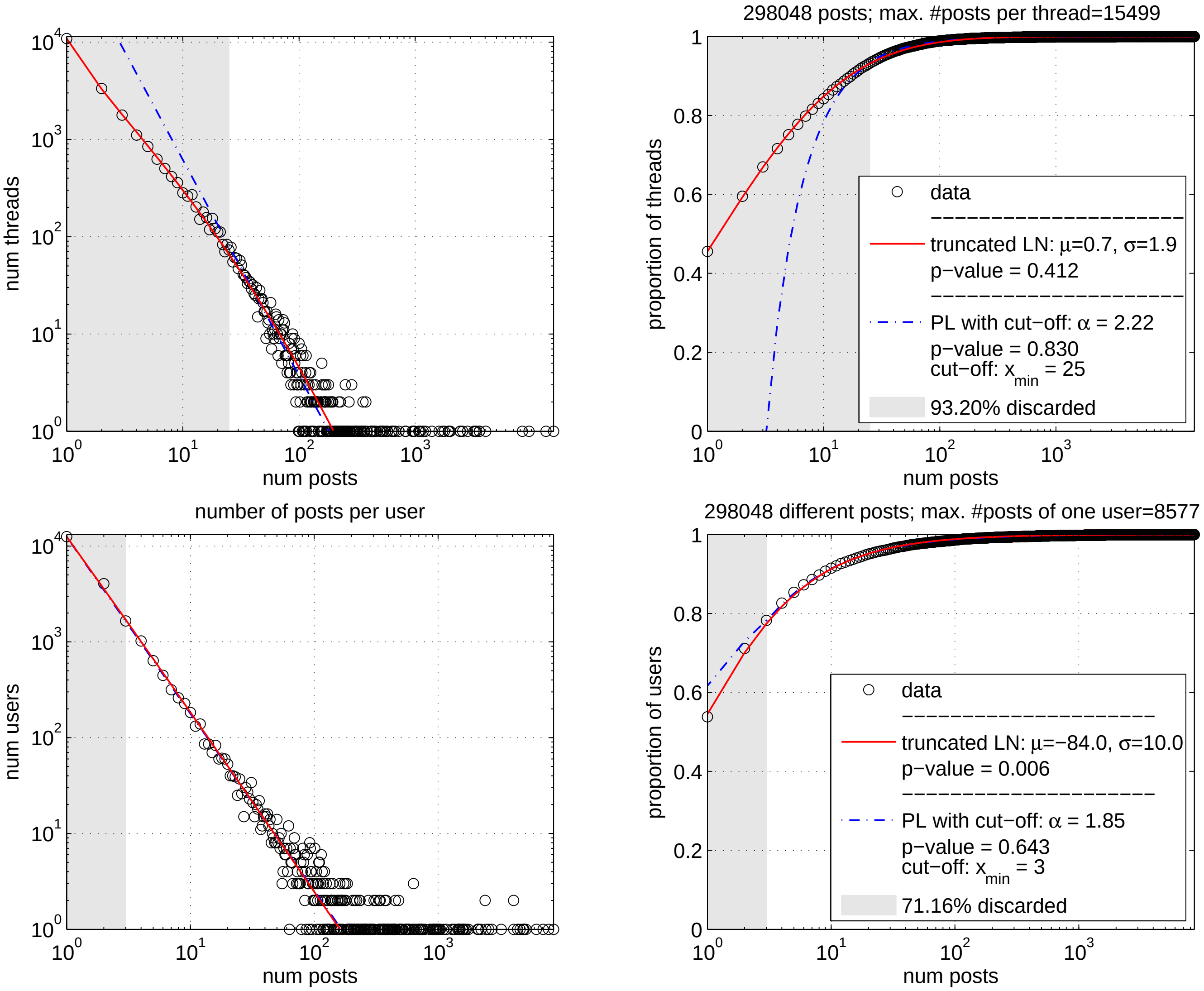

Figure 1: Heavy-tailed distributions of activity. The data (black circles) is fitted by a log-normal distribution (red curve) and a power-law (blue curve) with cut-off. White ares indicate the regions where the power-law should fit the data. Right sub-figures show the normalized cumulative version of the distributions on the left. Top: Distribution of the number of posts per thread. Bottom: Distribution of the number of posts per user.

When contrasting the percentage of threads which receive a certain amount of posts with the share of all posts within these threads, as is done in Table 3, we observe a variation of the typical 80/20 rule. In this case, 85% of the threads receive only 16% of the activity (posts) and the remaining 15% obtain 85% of the posts. And, to give a more striking example, the 0.1% of the threads that have received more than 1000 posts (around 250 threads) contain more than 30% of all activity (all posts) in the forums.

If we analyze the number of posts per user we obtain a similar picture, with the slight difference that this distribution resembles a nearly perfect power law (see Figure 1 bottom). Only users who write one or two posts (still 71.16% of all users) are not well approximated by a power law with an exponent of -1.85. The corresponding KS-test obtains a high a p-value of 0.63, while the approximation with log-normal distribution is rejected with a low p-value of 0.006. Again we observe a variation of the typical 80/20 rule (Table 4) when contrasting with the share of all posts. In this case, the more than 90.7% of all users which write less than 10 posts, produce less than 15% of the activity, and the remaining 9.3% is responsible for more than 85% of the entire amount of posts in the dataset. Even more surprising is the fact that only 47 users (0.2% of the total) produce more than 43% of all posts on MySpace in Spanish. We have thus a very small minority responsible for a huge part of the entire activity in the forums.

The difference in the distributions which fit best the data has an important impact on which type of model would be the most suitable to produce this heterogeneous behavior.

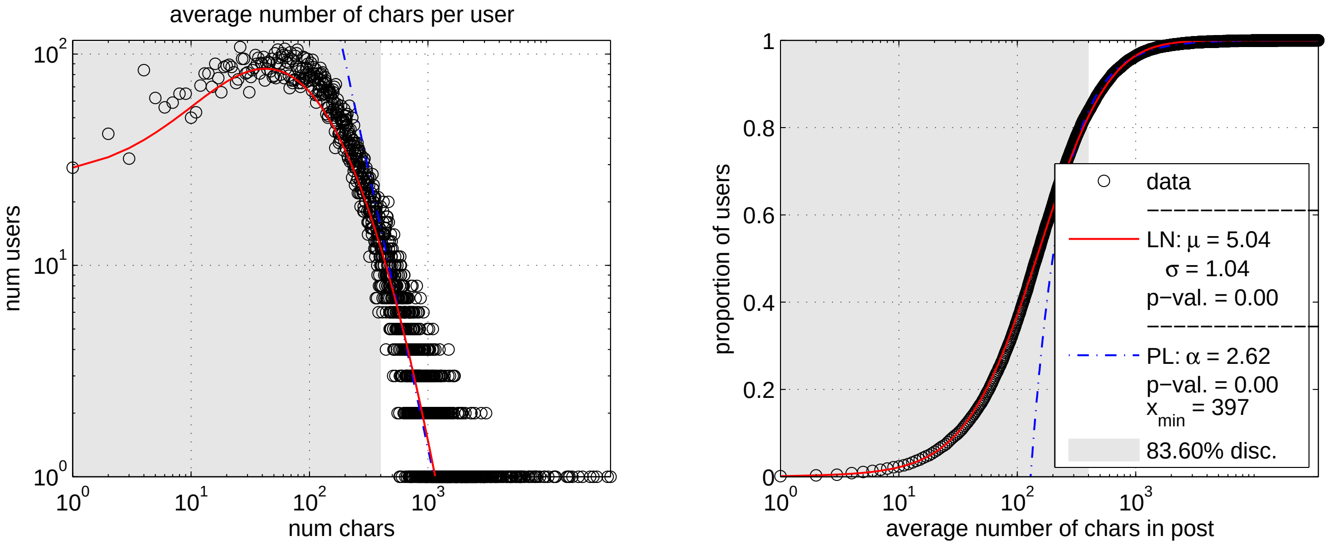

Figure 2: Distribution of the average length (number of characters) of the posts for every user. The data (black circles) is fitted by a log-normal distribution (red curve) and a power-law (blue curve) with cut-off. The white ares indicates the regions where the power-law should fit the data. Right sub-figure shows the normalized cumulative version of the distribution on the left.

# posts per thread	% of threads	% of posts
$[0, 1]$	47.77%	3.66%
$[2, 10]$	37.11%	12.52%
$[11, 100]$	14.02%	32.64%
$[101, 1000]$	0.99%	18.96%
$[1001, 10000]$	0.10%	22.55%
$[10001, 20000]$	0.01%	9.67%
$[0, 20000]$	100.00%	100.00%
$[0, 10]$	84.89%	16.17%
$[10, 20000]$	16.25%	84.77%
$[50, 20000]$	2.87%	61.49%

Table 3: Percentage of threads (second column) with a certain number of posts (first column) compared to their share of all posts (third column).

# posts per user	% of users	% of posts
$[1, 1]$	53.82%	4.21%
$[2, 10]$	37.68%	10.46%
$[11, 100]$	7.11%	15.99%
$[101, 1000]$	1.19%	26.18%
$[1001, 10000]$	0.20%	43.15%
$[1, 10000]$	100.00%	100.00%
$[1, 9]$	90.71%	14.06%
$[10, 10000]$	9.29%	85.94%
$[50, 10000]$	2.44%	75.08%

Table 4: Percentage of users (second column) with a certain number of posts (first column) compared to their share of all posts (third column).

Typical models which lead to a power law, such as preferential attachment [19], would have to be combined with a multiplicative model which produces a LN-distribution [17] to produce a process to account for the two types of heavy-tailed distributions we have described above.

3.3 Post length

Given this great amount of heterogeneity in the activity of users and in threads, one would expect the actual imprints of this activity (the posts) also to be of very different form and length. This is indeed the case when we compare the average lengths (i.e. its number of characters) of the posts of every user. Figure 2 shows the distribution of this quantity (bin-width=1), which looks at first sight quite exactly like a log-normal distribution. However, a KS-test rejects both the log-normal approximation of the entire distribution as well as the power-law fit of the tail. The rejection of the log-normal fit is caused by an outlier at an average post length of around four[2]. Nevertheless, to account for the log-normal shape of this distribution and as it is frequently done for LN-distributions [14] we will use its median and geometric standard deviation to compare in the next section different demographic groups. This helps to avoid excessive influence of users with very large average post lengths on the compared statistics (as it would be the case when using the mean for this type of data).

3.4 Demographic statistics

If we disclose the users participating in the forums by age, sex and origin we obtain the following conclusions.

The age-sex pyramid (Figure 3) shows a great heterogeneity in the structure of the population. Apart from the peaks at high ages, caused by people providing a wrong birth date

[2] Note that some posts only contain a link or an image. Such posts have a length of 0, since we only count the actual text characters when determining this quantity. It is therefore possible (and indeed the case) to have an average post length smaller than 1.

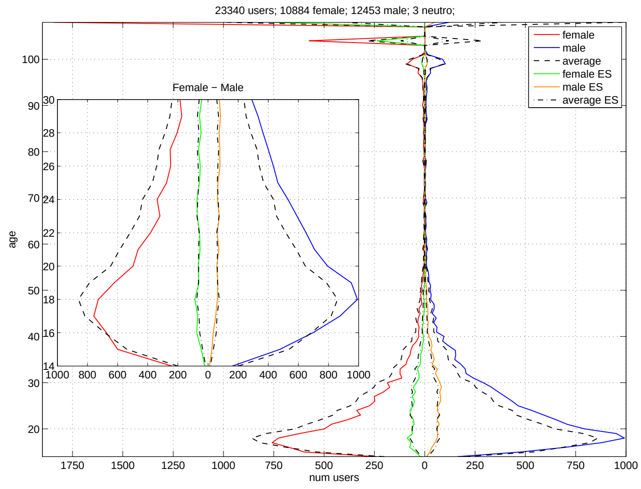

Figure 3: Population pyramid. We compare the entire Spanish speaking population of the MySpace forums with Spaniards among them. Inset shows an enlarged view of ages between 14 and 30. The Spaniards are nearly uniformly distributed in this age interval in contrast to the overall population.

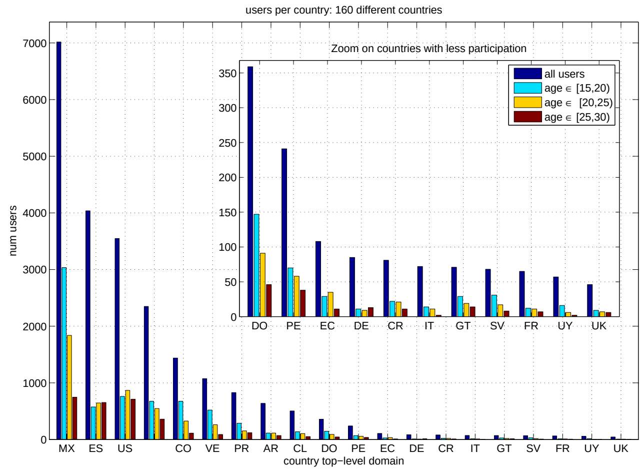

Figure 4: Distribution of the number of users per country in different age intervals confirms the different trends in population structure between Spain and Latin America. The fourth largest group of users has not released information on their home country. Inset shows countries with less users.

(probably the lowest possible at the moment of their registration) and more pronounced in the female users, the most frequent age is 17 years for female users (red curve) and 18 years for male users (blue curve) as shown in the inset on the left of Figure 3. The age structure between the sexes is almost balanced with a slight advantage for females among the youngest users but for users of 17 years and older the majority is male. This trend continues until ages older than 50 years where the number of users is too small to notice a significant difference. The median age is 22 for male users and 20 for females. Interestingly, this is a result contrary to [24] where a majority of female users was reported. This difference might either be a specialty of the Spanish speaking community in MySpace or the subset of users which participate in the forums. The masculine prevalence is even further pronounced when we consider the number of posts per sex. As shown in the top three rows in table 5 we found a slight majority of male users in our dataset, but this 53.4% majority generated 61.1% of all posts.

If we disclose these results further by country we obtain the following quite different pictures comparing for example the Spaniards among the MySpace users with those of other countries.

Contrary to the global average, the majority (56.9%) of the Spanish users is female and the structure of the ages (curves in green and orange in Figure 3) is fairly uniform among the users between 18 and 30 years, which results in a much larger median age (28 for the male and 24 for the female Spanish MySpace users). This structural difference between the users of different countries is also clearly visible in Figure 4, which shows the number of users of the most frequent countries of origin. The Mexican users represent the largest group (almost 30% of all users), and Spain obtains the second place in this ranking, with approximately 4000 users. In the same figure we have also added three bars disclosing the subset of users between 15 and 30 into three age intervals. We notice again that the Spaniards are divided fairly evenly into these age intervals, whereas Mexicans and other Latin American countries display a much more heterogeneous age structure, having more users between 15 and 20 (blue bars) than in the two reaming intervals together.

Curiously, U.S. American (and to a lower degree Argentinian) users show the same uniform age structure as Spaniards which might be caused by a combination of higher literacy rate of people over twenty with new technologies and a more uniform population pyramid in economically more developed countries. Colombia and Venezuela, the next ranked countries by their number of users (after the group of people who do not report their citizenship) show a similar structure as Mexico. The median age in these countries is also higher (25 for male and 24 for female US users), which is contrary to the fact that most of the 8 countries with the most users in our study show median ages of around 20 as shown in Table 5. The youngest users are found in Colombia with median of 19 (males) and 17 for females followed by Mexico, Chile and Venezuela.

When looking at the distribution of sexes five of the eight countries with the most users, presented in Table 5, exhibit a female majority. However, among the four countries with the highest number of users only Spain shows this majority, which explains the masculine prevalence in total. Interestingly, only in the case of Puerto Rico and Argentina the female users write more posts than the males and even more surprisingly their share of posts is always significantly smaller than their percentage (with the only exception of the US were it is slightly higher).

Finally we are also interested in who writes the longer posts. We have already seen that males on average seem to be more likely to post than females, but does the feminine poster compensate this by writing longer posts? Using the median of the average number of characters in the posts, we find that this is indeed the case. The average post of male

	male		female	
total	12453	53.4%	10884	46.6%
posts	182163	61.1%	115880	38.9%
age	med.=22	$\sigma = 9.7$	med.=20	$\sigma = 10.1$
length	med.=138.0	$\sigma_g = 3.4$	med.=153.8	$\sigma_g = 3.0$
ES	1737	43.0%	2299	56.9%
posts	28250	55.3%	22818	44.7%
age	med.=28	$\sigma = 10.4$	med.=24	$\sigma = 9.3$
length	med.=185.7	$\sigma_g = 3.1$	med.=183.7	$\sigma_g = 2.7$
MX	4073	58.0%	2944	42.0%
posts	80511	71.4%	32191	28.6%
age	med.=20	$\sigma = 8.4$	med.=19	$\sigma = 9.4$
length	med.=146.3	$\sigma_g = 3.1$	med.=154.0	$\sigma_g = 3.0$
US	2127	59.9%	1421	40.0%
posts	30830	57.6%	22691	42.4%
age	med.=25	$\sigma = 9.3$	med.=24	$\sigma = 11.4$
length	med.=95.3	$\sigma_g = 3.7$	med.=122.3	$\sigma_g = 3.5$
CO	839	58.3%	600	41.7%
posts	9032	66.6%	4527	33.4%
age	med.=19	$\sigma = 9.0$	med.=17	$\sigma = 9.6$
length	med.=170.0	$\sigma_g = 2.8$	med.=154.0	$\sigma_g = 2.7$
VE	530	49.3%	545	50.7%
posts	9811	59.0%	6818	41.0%
age	med.=22	$\sigma = 10.0$	med.=21	$\sigma = 9.9$
length	med.=151.2	$\sigma_g = 3.0$	med.=173.0	$\sigma_g = 3.1$
PR	332	40.0%	497	60.0%
posts	1779	43.1%	2348	56.9%
age	med.=22	$\sigma = 10.0$	med.=21	$\sigma = 9.9$
length	med.=103.1	$\sigma_g = 3.4$	med.=135.3	$\sigma_g = 3.3$
AR	262	41.0%	377	59.0%
posts	3107	49.9%	3123	50.1%
age	med.=26	$\sigma = 10.5$	med.=22	$\sigma = 8.4$
length	med.=179.7	$\sigma_g = 3.1$	med.=168.2	$\sigma_g = 2.7$
CL	206	40.8%	299	59.2%
posts	3854	63.0%	2264	37.0%
age	med.=22	$\sigma = 9.2$	med.=20	$\sigma = 9.7$
length	med.=138.0	$\sigma_g = 2.9$	med.=164.8	$\sigma_g = 2.6$

Table 5: Demographic statistics per home country (four rows per country). First row: Amount and percentage of male and female users. Second row: Amount and percentage of posts written by them. Third row: Median and stdv σ of their ages (only considering ages $<$ 90). Forth row: Median and geometric standard deviation σ_g of the number of chars in their posts. Top four rows correspond to the statistics of the entire dataset. σ and σ_g very similar among countries and sexes.

users is in 50% of the cases less than 138 characters long, while females write in the same proportion at least 153.8 chars on average. This would correspond to posts of female users about 3 or 4 words longer than those of their male colleagues.

When we disclose this result by the origin of the users, we find that Spaniards are the most chatty posters. Their median average post-length is about 48 (30 for females) characters longer than the one of the entire population. No significant difference can be found in this case between the sexes. There are only two countries of the ones analyzed in Table 5 where males write longer posts than females, namely Colombia and Argentina. The remaining countries follow the same trend as the totality of users. The two countries with the shortest posts are United States and Puerto Rico. Already the female users of these countries have a median average post length shorter than those of the males in the other six countries of Table 5 and their male compatriots top this by writing even shorter posts (by about 30 characters). It would be interesting to contrast these results with those of the MySpace forums in English, to check whether the close contact of the users with English in these two countries may cause this effect.

4. ANALYSIS OF SPECIFIC THREADS AND USERS

In this section we propose two new descriptive measures which allow to rank users by their amount of contributions to the different threads and to rank the threads themselves by the interest they provoked in the MySpace community. We use simple analytically calculated variables as measure for the ranking.

To find important discussions from bulletin boards a technique using automatic rule extraction and fuzzy decision trees has been proposed [20]. The main drawback of such methods is the necessity of large training sets to work properly and the dependency on the subjective criteria of who generated these training sets. Another related study trains a classifier using features that combine semantic information with structural characteristics such as the number of posts to measure the degree of controversy of a forum discussion [15]. Our proposed measure appears to be a more convenient indicator because of its simplicity, objectivity and robustness. It can be calculated efficiently and is monotonic (it never decreases), which makes it also a stable quantity to monitor and rank a discussion thread while it is still alive and receiving contributions. The same applies when we compare our measure for user ranking with related work using more complex network based ranking algorithms such as page rank to find users with high expertise in help-seeking forums [26].

4.1 Variable definition and operationalization

Both measure are adapted versions of the h-index [11], commonly used to characterize the scientific output of researchers. The papers of a researcher are ordered by their number of citations in descending order and the h-index is then defined as the maximum rank-number, for which the number of citations is greater or equal to the rank number. For example, if a scientist has an h-index of 11 it means that he has written 11 papers with at least 11 citations each. It represents a fair quantity which considers the number of papers published by the scientist and their visibility, or how often these papers are cited by other scientists. Some extensions of these index have been proposed as an alternative to the impact-factor of journals and conferences [4, 21]. See [3] for more details and a review on the literature about the h-index. Recently the h-index was used as well as a structural measure for nested conversations to account for the degree of controversy in them [10].

4.1.1 h-user index

To rank the users by their amount of contributions in the different threads we use a measure denominated h-user index. It is calculated for a single user by ordering the threads by the number of posts from this user in descending order.

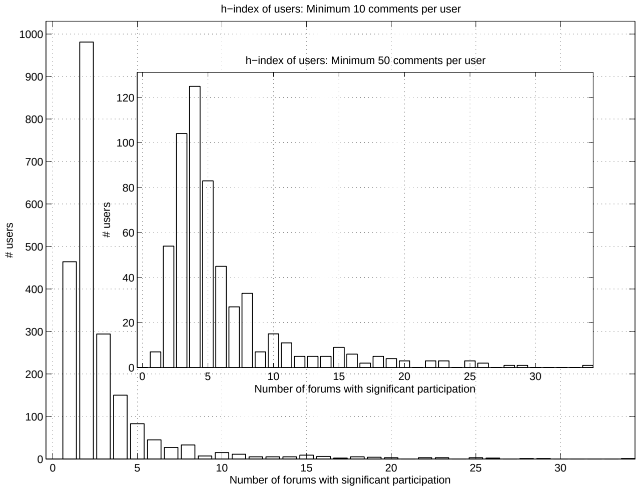

Figure 5: Distribution of the degree of contribution to the different threads (h-user index) for users who have posted at least 10 (or 50 in the inset) times.

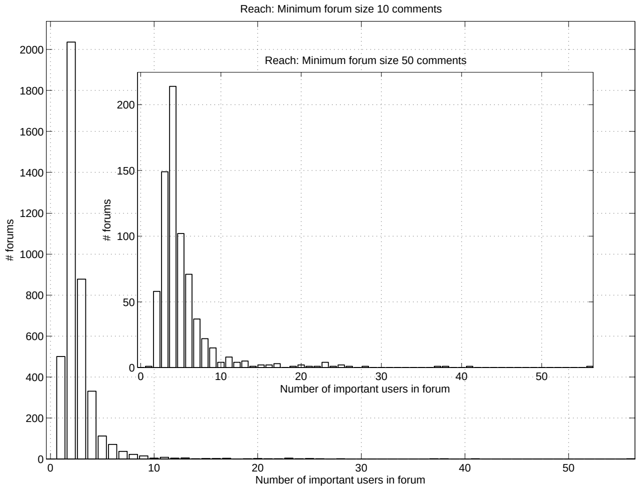

Figure 6: Distribution of the number of active users of threads with at least 10 (or 50 in detail) posts.

The h-user index is then the highest rank of a thread which fulfils that the number of posts of these users in the thread is higher than or equal to its rank number. For example, if a user posts in five different threads. In one 6 times, in the second and third 2 times and in the remaining two threads trice, we get the following sequence after ordering this post counts: $\{6, 3, 3, 2, 2\}$, which translates into an h-user index of 3 .

This measure has the advantage to measure the degree of participation in different threads, taking into account both the number of threads and the number of posts within these threads. A high number of posts or a participation in a large number of threads does not automatically correspond to a high h-user index. Numbers such as the average amount of posts per thread cannot account for this relation.

For instance, for a user who participates only in a single thread has an h-user index of 1, no matter how many times he posts under this thread. Likewise, a user who has posted in more than 100 different threads would as well achieve only an h-user index of 1 if he only posted exactly once in all of these threads. In this case he does not contribute more than the user who only posted in a single thread to the elaboration of the conversation. The user needs to engage more and write several posts in the same threads to achieve a greater h-user index.

Figure 5 shows the distribution of the h-user indexes of all users who write more than 10 (or 50 in the inset) posts. We notice that the vast majority of users only contribute to very limited degree to the elaboration of the conversations. The distribution peaks at an h-user index of 2 if we consider only the users who post 10 or more times and at 4 if we cut of users with less than 50 posts. Only a very small number of users contributes in a way to achieve an h-user index larger than 10.

4.1.2 *h-thread index*

If on the other hand we would like to measure the relevance of a thread by the amount of different users who participate actively in it, we can use the h-thread index which we calculate in the following way:

For a single thread we order the users who post in this thread by their amount of posts. The thread's h-index (or h-thread index) is then defined as the highest user rank, who fulfils that the user posted more or equally often in this thread than its rank number. For example we have a thread where 4 users post about a certain topic. UserA posts 7 and UserB 9 times, UserC only once and UserD twice, which results in an ordered sequence of $\{9, 7, 2, 1\}$ and an h-thread index of 2, meaning that in this case only two users run the interaction. The remaining two participate only marginally and probably do not contribute really to the ongoing conversation in the thread.

To achieve a global vision about this variable we show in Figure 6 the h-thread index of all threads that have obtained more than 10 (or 50 in the inset) posts. We observe that the vast majority of the threads has less than 10 users with a relevant participation. The peak is at around 4 users for threads with more than 50 posts and 2 for threads with more than 10 posts. Which implies that most threads only involve a very reduced number of users who run the interaction.

4.2 Rankings

In the following we present two tables with the top 20 threads and users according to the two variants of the h-index defined in the previous section.

4.2.1 *Ranking of threads*

First we focus on the threads. Table 6 shows the h-thread index as well as some other descriptive variables like the number of posts and users as well as the median number of characters of the posts for the 20 threads with the highest h-thread index. Within brackets we can find the rank obtained when ordering by one of these other variables.

Interestingly the ranks of the median length of the posts are very high indicating that if the users participate several times in a the same thread they tend to write shorter contributions.

Nr.	h	posts-Tot.	(#)	user-Tot.	(#)	post-length	(#)	thread-name
1	56	15499	(1)	277	(2)	64	(17935)	palabras encadenadas
2	41	8358	(4)	200	(10)	34	(21122)	ke piensas de la persona de arriba.....
3	38	9540	(3)	247	(4)	7	(23238)	tu ultimo pensamiento
4	37	13327	(2)	113	(39)	44.5	(20060)	raptor n kike's bar
5	28	3566	(7)	128	(31)	40	(20523)	donde invitaras a salir a...
6	26	2822	(11)	230	(6)	589	(1955)	di una mentira
7	25	3275	(9)	164	(15)	46	(19867)	qué ests haciendo?
8	25	2433	(14)	207	(7)	34	(21123)	está bonita la persona arriba de ti?
9	24	2559	(12)	90	(62)	43	(20174)	el ultimo es postear gana
10	23	3354	(8)	157	(18)	456	(2799)	solo para feos
11	23	3198	(10)	117	(37)	47	(19767)	juguego de preguntas
12	23	2005	(15)	132	(29)	4	(23362)	palabras que empiecen por a
13	23	1961	(16)	204	(9)	67	(17578)	que cancion estas escuchando ahora
14	22	4045	(5)	102	(48)	44	(20081)	cuenta regresiva (5000 al 0)
15	21	1783	(18)	140	(25)	4	(23363)	califica a la foto de la ...
16	20	1927	(17)	81	(73)	25	(21982)	matando el aburrimiento
17	20	1226	(22)	139	(26)	114	(13191)	*** pasando lista en el foro*
18	19	1073	(26)	109	(41)	139	(11214)	elige
19	17	3569	(6)	238	(5)	319.5	(4472)	contando asta el 5 mil
20	17	1723	(19)	112	(40)	57	(18676)	jugemos verdad o reto te atreves a jugar

Table 6: Statistics and ranking of the 20 threads with the highest h-index, i.e. number of important users (first column). Within brackets the rank obtained when ordering the threads by other variables. Column 2: Total number of posts in the thread. Column 3: Number of unique users who participate in the thread. Column 4: Median of the length of the posts (in chars) in the thread.

Nr.	h	posts	(#)	threads	(#)	initiate	(#)	post-length	(#)	S	Age	State	Username
1	34	7823	(2)	1138	(3)	143	(4)	185.0	(9320)	M	21	US	UserA
2	29	7066	(3)	349	(15)	31	(50)	40.1	(19940)	M	39	ES	UserB
3	28	8577	(1)	374	(11)	12	(218)	48.8	(19212)	F	18	MX	UserC
4	26	5136	(5)	307	(21)	45	(21)	54.4	(18690)	F	30	US	UserD
5	26	4950	(6)	625	(4)	35	(38)	150.3	(11169)	M	23	MX	UserE
6	25	4905	(7)	280	(23)	29	(58)	61.0	(18117)	M	19	VE	UserF
7	25	4855	(8)	609	(5)	28	(61)	125.1	(12823)	M	27	MX	UserG
8	25	4084	(13)	536	(7)	24	(75)	131.2	(12388)	M	28	MX	UserH
9	23	4607	(9)	343	(17)	32	(46)	68.1	(17438)	F	28	HN	UserI
10	23	4401	(10)	1428	(2)	725	(1)	298.1	(5458)	M	24	MX	UserJ
11	23	3242	(14)	182	(56)	26	(68)	77.3	(16667)	F	22	US	UserK
12	22	6228	(4)	425	(9)	14	(152)	110.4	(13924)	M	24	MX	UserL
13	22	2571	(16)	279	(25)	24	(76)	56.8	(18495)	M	27	MX	UserM
14	22	2404	(19)	279	(24)	67	(12)	41.2	(19840)	F	25	US	UserN
15	20	2128	(21)	402	(10)	35	(41)	62.6	(17937)	F	25	ES	UserO
16	20	1690	(25)	190	(51)	40	(28)	68.7	(17419)	M	18	US	UserP
17	20	1573	(29)	209	(41)	37	(32)	134.6	(12164)	F	18	US	UserQ
18	19	4084	(12)	586	(6)	18	(106)	115.6	(13550)	M	25	MX	UserR
19	19	2725	(15)	323	(18)	40	(29)	106.2	(14255)	F	25	NL	UserS
20	19	1594	(28)	187	(55)	4	(893)	425.5	(3426)	F	16	MX	UserT

Table 7: Statistics and ranking of the 20 users with the highest h-index. The numbers within brackets show the ranks obtained if all users were all ordered by the amount of the corresponding column. Column 1: Number of threads with a significant participation of the user (h-user index). Column 2: Total number of posts of the user. Column 3: Number of different threads where the user participated. Column 4: Number of threads initiated by the user. Column 5: Average length (number of chars) of the posts. Column 6: Sex of the user: (M)ale or (F)emale. Column 7: Age of the user. Column 8: Country of origin.

In case where two threads achieve the same h-thread index we use the number of posts to break the tie. The most successful thread, entitled "palabras encadenadas" ("word chain" in English) achieves an h-thread index of 56 and is also ranked first when ordered by the number of posts. The objective of this thread is simple to build a long chain of words. The users have to post a word starting with the last syllable of the word in the previous post. This thread is somehow symptomatic for the type thread which achieves the greatest amount of participation. Often the initiator posts a question which incites the users to respond to something written in the previous post. Some threads are however even simpler, e.g. the threads “contando asta el 5 mil” (“counting to 5000”), where the users simple try to reach 5000 posts or “el ultimo es postear gana” (“the last post wins”)[3], a long chain of posts claiming everyone being the last one to post in the thread and therefore the winner. In this case it does not seem to be adequate to speak about conversation, the threads are more some kind of pastime, with little actual content.

A noteworthy exception seems to be the thread “raptor n kike’s bar” (ranked number 4 in table 6) where a reduced number of users considering themselves as the “owners” of the MySpace-forums run a virtual bar and serve drinks only to members of their virtual tribe. The members of this tribe also appear in a prominent position in the ranking of users as we will see in the next section. Their behavior and the relation among them has been investigated by the means of virtual ethnography [2].

4.2.2 Ranking of users

In Table 7 we show the top 20 users ranked by their h-user index. To solve the user’s privacy we changed their names to a simple combination of User plus a letter. The table presents apart from their h-user index as well information about their number of posts, the amount of different threads the users have participated in, how many threads they have initiated, the average length of their posts and data like age, sex and country of origin. To break the ties in the ranking we use again as in the case of the threads the number of posts.

The top ranked user achieves an h-user index of 34 and is ranked second by the number of posts and forth by the number of different threads where he has participated. Interestingly, the top ranked user in the later category only achieves an h-index of 18 (not show in table 7) and is ranked thus only number 23 by its h-user index. We also observe a considerable drop in the rank of UserL, ranked forth by his number of posts and ninth by the number of threads, he only achieves an h-index of 22, ranking him as number twelve in table 7. UserK, UserP, UserQ and UserT on the other hand gain positions achieving a much higher rank with their h-user index than according to the other variables shown in this ranking.

Again we observe the low ranks according to the length of the posts of these users, confirming that writing more posts implies writing shorter ones.

5. CONCLUSIONS AND DISCUSSION

In the first part of this study we have provided an in depth analysis of the MySpace forums in Spanish. We present statistics for the Spanish language usage and found highly heterogeneous posting activity. Most of the users and threads write and receive only a reduced number of posts (nearly 50% only 1) and only a small number of users and threads creates the mayor amount of activity. These heterogeneous distributions of activity in the form of heavy tailed distributions have also been observed in similar settings such as the number of articles submitted to a bulletin board system [18], the number of comments to Slashdot stories [12] or the number of messages in a friendship network [9]. The combination of a log-normal distribution for the number of posts per thread and the more powerlaw like shape of the distribution of the number of posts per user, demands a model being able to express both types of behavior. Such a model might serve as a bridge between competing models creating either log-normal or power-law distributions to explain a certain dataset [17] and is subject of current research.

When analyzing the demographic structure of the users in the forums we found interesting differences in the age structure among different countries. Countries more developed economically have a more uniform age distribution among the users from 15 to 30 while users from other countries show a decaying age distribution among these ages. The overall user structure shows a majority of male posters in the forums, which is different from what is reported in a recent study of MySpace user profiles [24], where a majority of female MySpace users is reported. To answer the question whether this difference is caused by specialties of the Spanish speaking community or is typical for posts in forums we will perform a similar study in the future with the MySpace forums in English. Furthermore, it would be interesting to contrast the findings with those from other social network sites, statistics of Internet usage and overall population structure, to get insight about the generality of this findings. Independently from the outcome of such a study we can conclude that the differences found in the demographic structure among different countries are of great importance for online marketing, the potential public for an advertising campaign would be very different if the target would be the Spanish or the Mexican market, for example.

Apart from an unequal population structure we also find different posting behavior among the sexes. Males write on average more posts than females, but the length of those messages is shorter than those of the feminine posters. Those differences may vary according to the home country of the users but should be considered and further investigated in future studies about user behavior in social networking sites.

Finally we would like to clarify that we do not have access to data (log-files) to assess how many people actually read the MySpace forums, so nothing can be said about the actual impact of these forums. We rather present a description of the subset of Spanish speaking users who actually participate in them. Its size is of the same order of magnitude as the one analyzed in [24].

In the second part of this article we introduce two variants of the h-index to rank the users of the forums by their amount of contributions in different threads and the threads by the amount of users who participate actively in a thread. The h-user index provides a nonlinear combination of the total number of posts and the amount of different threads where a user is active, while the h-thread index combines the number of unique users in a thread with the number of posts it receives. The resulting indexes are monotonic in time, robust and easy to calculate and can be applied on other type of forums as well. If the forums allow nested posts, they can be combined with a third variant involving an h-index of the structure of the conversations as proposed in [10].

The rankings have been applied to find and select important users and threads for a study on virtual ethnography in the MySpace forums [2]. However, a more detailed study is needed to contrast our proposed ranking algorithms with those of users manually inspecting the threads, machine learning algorithms for important thread detection [20] and network based ranking algorithms for users [26].

[3] Note that the incorrect Spanish spellings of the two last examples are literal copies from the website.

6. ACKNOWLEDGMENTS

We thank MediaPro for the permission to publish this work and Mariona Estrada, Raquel Navarro, Pablo Rebaque and Irene Sancho for helping to elaborate this manuscript. This work has been developed inside the i3media project (CDTI 2007-2010) and partially funded by the Centro para el Desarrollo Tecnológico Industrial (CDTI), within the Ingenio 2010 initiative and the Càtedra Telefónica de Producció Multimèdia.

7. REFERENCES

[1] V. Adamchik and H. Srivastava. Some series of the zeta and related functions. *Analysis*, 18:131–144, 1998.

[2] E. Bondia, A. Kaltenbrunner, P. Rebaque, M. Estrada, I. Sancho, R. Banchs, and R. Navarro. Detección de relaciones entre usuarios, identificación de clases de usuarios e identificación de hubs. Technical Report A10T1H2S2 (i3media internal), Barcelona Media, 2008.

[3] L. Bornmann and H. D. Daniel. What do we know about the h index? *J. Am. Soc. Inf. Sci. Tech.*, 58:1381–1385, 2007.

[4] T. Braun, W. Glanzel, and A. Schubert. A Hirsch-type index for journals. *Scientist*, 19:8–8, 2005.

[5] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 57–70, New York, NY, USA, 2008. ACM.

[6] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. e-print arXiv:0706.1062, 2007.

[7] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, New York, NY, USA, 3rd edition, 2002.

[8] A. Fumero and J. M. García Hervás. Social networks. contextualizing the phenomenon of web 2.0. *TELOS. Cuadernos de Comunicación e Innovación*, (76), 2008. (in Spanish, online).

[9] S. A. Golder, D. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *3rd International Conference on Communities and Technologies (CT2007).*, 2007.

[10] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 645–654, New York, NY, USA, 2008. ACM.

[11] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569, 2005.

[12] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López. Homogeneous temporal activity patterns in a large online communication space. *IADIS International Journal on WWW/INTERNET*, 6(1):61–76, 2008.

[13] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[14] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *Bioscience*, 51:341–352, 2001.

[15] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *WWW2006, 3rd Annual Workshop on the Weblogging Ecosystem*, Edinburgh, UK, 2006.

[16] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[17] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[18] K. Naruse and M. Kubo. Lognormal distribution of bbs articles and its social and generative mechanism. In *WI'06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 103–112, Washington, DC, USA, 2006. IEEE Computer Society.

[19] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[20] S. Sakurai and R. Orihara. Discovery of important threads from bulletin board sites. *Int. J. of Information Technology and Intelligent Computing*, 1(1):217–228, 2006.

[21] A. Sidiropoulos and Y. Manolopoulos. Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, 79(12):1679–1700, 2006.

[22] K. Sigman. Appendix: A primer on heavy-tailed distributions. *Queueing Systems*, 33:261–275, 1999.

[23] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 655–664, New York, NY, USA, 2008. ACM.

[24] M. Thelwall. Social networks, gender, and friending: An analysis of myspace member profiles. *J. Am. Soc. Inf. Sci. Technol.*, 59(8):1321–1330, 2008.

[25] Universal McCann. Internacional Social Media Rearch. Wave 3. published online, March 2008. `http://www.universalmccann.com/Assets/wave_3_20080403093750.pdf`.

[26] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.