An Experimental Analysis of Suggestions in Collaborative Tagging

Dirk Bollen *
* *Faculty of Innovation Sciences and Industrial Engineering*
University of Technology Eindhoven
Eindhoven, The Netherlands
Email: d.g.f.m.bollen@tue.nl

Harry Halpin †
† *School of Informatics*
University of Edinburgh
Edinburgh, United Kingdom
Email: H.Halpin@ed.ac.uk

***Abstract*—Most tagging systems support the user in the tag selection process by providing tag suggestions, or recommendations, based on a popularity measurement of tags other users provided when tagging the same resource, like a web-page. In this paper we investigate the influence of tag suggestions on the emergence of power-law distributions as a result of collaborative tag behavior. Although previous research has already shown that power-laws emerge in tagging systems, the cause of why power-law distributions emerge is not understood empirically. The majority of theories and mathematical models of tagging found in the literature assume that the emergence of power-laws in tagging systems is mainly driven by the imitation behavior of users when observing tag suggestions provided by the user interface of the tagging system. This imitation behavior leads to a feedback loop in which some tags are reinforced and get more popular which is also known as the 'rich get richer' or a preferential attachment model. We present experimental results that show that the power-law distribution forms when tag suggestions are not presented to the users, and the power-law distribution does not hold when there are tag suggestions presented to the user. Furthermore, we show that the real effect of tag suggestions is rather subtle; the power-law distribution that would naturally occur without tag suggestions is 'compressed' if tag suggestions are given to the user, resulting in a shorter long tail and a 'compressed' top of the power-law distribution. The consequences of this experiment show that tag suggestions by themselves do not account for the formation of power-law distributions in tagging systems.**

***Keywords*-Distributed information systems; Information retrieval; User interfaces**

I. Introduction

During the last decade the Web has become a space where increasing numbers of users create, share and store content, leading it to be viewed not only as an "information space" [1] but also a "social space" [2]. This new step in the evolution of the Web, often referred to as the "Web 2.0," was shaped by the arrival of the different services that came into existence to support users to easily publish content on the Web, such as photos (Flickr), bookmarks (del.icio.us), movies (YouTube), blogging (Wordpress), and so on [3]. Almost simultaneously with the growth of user-generated content on the Web came a need create order in this fast growing unstructured data. Tagging has become the predominant method for organizing, searching and browsing online web-pages, as well as any other resource.[1] Tagging refers to the labeling of resources by means of free-form descriptive keywords. With tagging users themselves annotate resources by tags they freely chose and thus forms a 'flat space of names' without the predefined and hierarchical structure characteristic of classic 'ontologies' in knowledge engineering.

Empirical studies of del.icio.us show that the number of tags needed to describe a resource consistently converges to a power-law distribution as a function of how many tags it receives [4]. We refer to the highest ranked frequencies of the power-law distribution as the 'top' of the distribution, as opposed to the long tail. Furthermore, we can consider the formation of a power-law distribution to be 'stable' known as *scale invariance*. A power-law distribution produced by tagging is a good sign of stability since, due to scale invariance, increasing the number of tagging instances only proportionally increases the scale of the power-law, but does not change the parameters of the power-law distribution. Thus, the first step in determining if users have reached a stable consensus in tagging is the detection of a power-law distribution from the frequencies of tags [5]. The reasons behind the emergence of a power-law distribution in tagging systems are yet unknown, although explanations fall into two general categories. The first of these explanations is relatively simple: the tags stabilize into a power-law because users are imitating each other via tag suggestions put forward by the tagging system [4]. The second and more recent explanation is that in addition to imitation, the users share the same background knowledge [6]. However, drawing these two influences apart has not yet been tested scientifically. We will proceed to attempt do this after reviewing in detail the various explanations of the emergence of power-laws in tagging.

II. Models of collaborative tag behavior

A. Formalizing Tagging

The traditional tripartite model of tagging is well-known. In essence, in a *tagging instance* a user u applies n tags $(t_1 ... t_n)$ in order to categorize a given resource r. There are

[1] A resource is anything that can be given a URI (Uniform Resource Identifier, including but not limited to web-pages [1].

three metrics that are often used to describe tagging systems. The first is the *tag-resource distribution*, which inspects the frequency that each tag $t_1....t_n$ has been applied to a given resource (such as a web-page) r by a number of distinct users $u_1...u_x$. In general, when we are referring to a distribution we are referring to the tag-resource distribution. This distribution is graphed by ordering the tags $t_1...t_k$ in descending rank order on the x axis against their frequency on the y axis. Further metrics that are of interest to researchers are *tag-growth distributions*, which counts the number of distinct tag assignments over some period of time over all users and resources in a tagging system. Another distribution is the *tag-correlation distributions*, which is the tag frequency for two tags t_i and t_j occurring in the same tagging instance.

B. A simple model: The Polya Urn

The most elementary model of how a user selects tags when annotating a resource is simple imitation of other users. Note that 'imitation' in tagging systems means that the tags are being reinforced via a 'tag suggestion' mechanism, and so the terms "imitation", "reinforcement", "feedback", and 'tag suggestion' can be considered to be synonymous in the context of tagging systems. The user can imitate other users precisely because the tagging systems tries to support the user in the tag selection process by providing tag suggestions based on tags other people used when tagging the same resource. There are minor variants of this theme, such as the possibility of using a combination of tags of other users in combination with a user's own previously used tags. In most tagging systems like del.icio.us these tag suggestions are presented as a list of tags that the user can select in order to add them to their tagging instance. The selections of tags from the tag recommendation forms a positive feedback loop in which more frequent tags are being reinforced, thus causing an increase in their popularity, which in turn causes them to be reinforced further and exposed to ever greater numbers of users. This simple type of explanation is easily amendable to preferential attachment models, also known as 'rich get richer' explanations, which are well-known to produce power-law distributions. Intuitively, the earliest studies of tagging observed that users imitate other pre-existing tags [4]. Golder and Huberman proposed that the simplest model that results in a "power-law" would be the classical Polya urn model [4]. Imagine that there is urn containing balls, each of some finite number of colors. At every time-step, a ball is chosen at random. Once a ball is chosen, it is put back in the urn along with another ball of the same color, which formalizes the process of feedback given by tag suggestions. As put by Golder and Huberman, "replacement of a ball with another ball of the same color can be seen as a kind of imitation" where each color of a ball is made equal to a natural language tag and since "the interface through which users add bookmarks shows users the tags most commonly used by others who bookmarked that URL already; users can easily select those tags for use in their own bookmarks, thus imitating the choices of previous users" [4]. Yet, this model is too limited to describe tagging, as it features only reinforcement of existing tags, not the addition of *new* tags.

C. Imitation and The Yule-Simon Model

The first model that formalized the notion of new tags was proposed by Cattuto et al. [7]. In order for new tags to be added, a single parameter p must be added to the model, which represents the probability of a new tag being added, with the probability $\bar{p} = (1 - p)$ that an already-existing tag is reinforced by random uniform choice over all already-existing tags. This results in a Yule-Simon model, a model first employed by Yule [8] to model biological genera and later Simon to model the construction of a text as a stream of words [9]. This model has been shown to be equivalent to the famous Barabasi and Albert algorithm for growing networks [10]. Yet the standard Yule-Simon process does not model vocabulary growth in tagging systems very well, as noticed by Cattuto et al. as it produces exponents "lower than the exponents we observe in actual data" [7].

Cattuto et al. hypothesize that this is because the Yule-Simon model assumes users are choosing to reinforce ($\bar{p}$) tags uniformly from a distribution of *all* tags that have been used previously, so Cattuto concludes that "it seems more realistic to assume that users tend to apply recently added tags more frequently than old ones" [7]. This behavior could be caused by the exposure of a user to a feedback mechanism, such as del.icio.us tag suggestion system. This suggestions exposes the user only to a subset of previously existing tags, such as those most recently added. Since the tag suggestion mechanism only encourages more recently-added tags to be re-enforced with a higher probability, Cattuto et al. added a memory kernel with a power-law exponent to standard Yule-Simon model. This means that the weight of a previously existing tag being reinforced is weighted according to a power-law itself, so that a tag that has been applied x steps in the past is chosen with a probability $\bar{p}(x) = a(t)/(x+\tau)$, where $a(t)$ is a normalization factor and τ "is a characteristic time scale over which recently added words have comparable probabilities" [7]. While the parameter p controls the probability of reinforcing an existing tag, this second parameter τ, controls how fast the memory kernel decays and so over what time-scale a tag may likely count as 'new' and so be more likely to be reinforced. As Cattuto et al. notes, "the average user is exposed to a few roughly equivalent top-ranked tags and this is translated mathematically into a low-rank cutoff of the power-law" [7]. This model produces an "excellent agreement" with the results of tag-correlation graphs [7]. It should be clear that the original Yule-Simon model simply parametrizes the probability of the imitation of existing tags. The modified Yule-Simon model with a power-law memory kernel also depends on the imitation of

existing tags, where the probability of a previously-used tag is decaying according to a power-law function.

D. Adding Parameters and Background Knowledge

Although Cattuto et al.'s model is without a doubt an elegant minimal model that captures tag-correlation distributions well, it was not tested against tag-resource distributions [7]. Furthermore, as noticed by Dellschaft and Staab, Cattuto et al.'s model also does not explain the sub-linear tag vocabulary growth of a tagging system [6]. Dellschaft and Staab propose an alternative model, which adds a number of new parameters that fit the data produced by tag-growth distributions and tag-resource distributions better than Cattuto et al.'s model [6]. The main points of interest in their model is that instead of a new tag being chosen uniformly, the new tag is chosen from a power-law distribution that is meant to approximate "background knowledge." So besides "background knowledge" ($\bar{p}$), their model also features the inverse of "background knowledge," i.e. the "probability that a user imitates a previous tag assignment" (p) [6]. In essence, Dellschaft and Staab have added (at least) two new parameters to a Yule-Simon process, and these additional parameters allows the reinforcement of existing tags to be more finely tuned. Instead of a single power-law memory kernel with a single parameter τ, these additional parameters allow the modeling of "an effect that is comparable to the fat-tailed access of the Yule-Simon model with memory" while keeping tag-growth sub-linear [6]. The model proposed by Cattuto et al. kept the tag-growth parameter equal to 1 and so makes tag growth linear to p [7]. Yet for us, most important advantage of Dellschaft and Staab over Cattuto et al.'s model is that their added parameters lets their model match the previously unmatched observation by Halpin et al. of the frequency rank distribution of resources being a power-law [5]. The match is not as close as the match with vocabulary growth and tag correlations, as resource-tag frequency distributions vary highly per resource, with the exception of the drop in slope around rank 7-10 [5].

E. Research Questions

What unifies all of these models is that they assume that imitation, usually assumed to be tag suggestions from the tagging system, has a major impact on the emergence of a power-law distribution. With concern to the modified Yule-Simon model and the more highly parametrized model that takes into account 'background knowledge,' different claims are made of where the imitated tags come from. Cattuto et al. proposes that they come from a random uniform distribution of tags while Dellschaft and Staab propose a more topic-related distribution that itself has a power-law distribution [6]. However, just because a simple model based on imitation of tag suggestions can lead to a power-law distribution does not necessarily mean that tag suggestions are actually the causal mechanism that causes the power-law distribution to arise in tagging systems. The research question posed then is: Is the tag suggestion mechanism, the main force behind the observed power-law distributions in tagging systems?

III. Experimental Design

In order to measure the effects of tag suggestions on the tag behavior of users we developed a Web-based experiment in which users were asked to tag 11 websites, with two varying conditions: the 'tag suggestion' condition (Condition A) in which 7 tag suggestions were presented to the user, and a 'no tag suggestion' condition (Condition B) in which no tag suggestions were presented to the user.

In this experiment we focus on del.icio.us which is the one of the earliest and well-known tagging systems. Del.icio.us was the first to introduce a tag based collaborative bookmark system. Del.icio.us has more than five million users and 150 million tagged URIs and so provides a vast data-set. The user interface used in our experiment presented the tag suggestions in a similar way to del.icio.us to avoid confusion.

The 11 websites used in the experiment were selected according to two criteria. First, the topics of the websites needed to appeal to the general public. Second, the website needed to have over 200 tagging instances. The appeal to the general public was operationalized by randomly choosing sites that were tagged with the tag "lifestyle" on del.icio.us. The tag "lifestyle" is a popular tag with 72,889 tagged web-pages as of October 2008. This was chosen in order to not bias our study to one particular specialized subject matter, and so exclude web-pages on del.icio.us that have a highly technical content. Specialized content may not lead to normal tagging behavior from users in the experiment who might not be familiar with the specialist subject matter. The second criteria of using only web-pages with over 200 tagging instances was chosen since it has been shown that stable power-law tag distributions emerge around the 100-150th tagging [4]. We did not want the tag suggestions to be from non-stable tag distributions, as it has been shown that the variance between the top popular tag could vary widely before 100-150th tag. The 11 websites selected for this experiment, with the popular tags provided from del.icio.us and the number tags. Note that while the number of URIs 11 may appear to be small, it is larger than previous experiments over tag suggestions [11] and was enough to give the experiment enough power to be statistically significant. It was far more critical for this experiment to get enough subjects in order for power-law distributions to be given the chance to arise without tag suggestion, and this would require at least 100 experimental subjects tagging each URI.

Figure 1 shows the experimental design. In the 'no tag suggestion' condition (Condition A), as shown in Figure 1, a user is presented the 11 websites he needs to tag without any

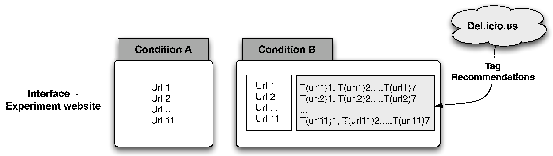

Figure 1. Experimental Design

form of tag suggestions. In the 'tag suggestion' condition (Condition B), also shown in Figure 1, a user is presented the 11 websites with 7 suggested tags. While the details of the tag suggestion algorithm applied by del.icio.us is unknown, for our experiment the suggested tags in condition B were aggregated from del.icio.us and the 7 suggested tags given by del.icio.us for each of the 11 websites. For the experiment the 7 popular tags were aggregated and presented to the participants in manner similar to how tags are suggested to users of del.icio.us, being shown to the user before they commence their tagging. Each of the 300 participants was randomly assigned to either the 'tag suggestion' or 'no tag suggestion' condition. Of these 300 users, 78 did not tag any website (37 in the 'tag suggestion' condition, 41 in the 'tag suggestion' condition) and are therefore excluded from further analysis. The users were randomized over age, gender, computer, Internet and their past tagging usage.

IV. Results

In total the 222 participants applied 7,250 tags over all websites in both conditions, with 3,694 tags applied in the 'tag suggestion' condition and 3,556 in the 'no tag suggestion' condition. On average every user in the 'tag suggestion' condition applied 32.69 ($S.D. = 9.77$) tags over all 11 URIs and for the no tag suggestion conditions 32.61 ($S.D. = 6.80$) tags over 11 URIs.

A. Detecting Power-Law Distributions

The power-law distribution is defined by the function:

$$y = cx^{-\alpha} + b \tag{1}$$

in which c and α are the constants that characterize the power-law and b being some constant or variable dependent on x that becomes constant asymptotically. The α exponent is the scaling exponent that determines the slope of the distribution before the long tail behavior begins. A power-law function can be transformed to a log-log scale as in the following equation:

$$log(y) = -\alpha log(x) + log(c) \tag{2}$$

This equation shows the characteristic property of power-law function is that when transformed to a log-log scale the power-law distribution takes the shape of a linear function with slope α. So transforming a function to a log-log scale and determining the slope α is one of the first steps in examining whether a distribution has a power-law. We averaged the tag-resource distributions for all 11 web-pages, and this distribution in log-log space is given in Figure 2. In a log-log scale, *both* conditions appear visually to exhibit power-law behavior.

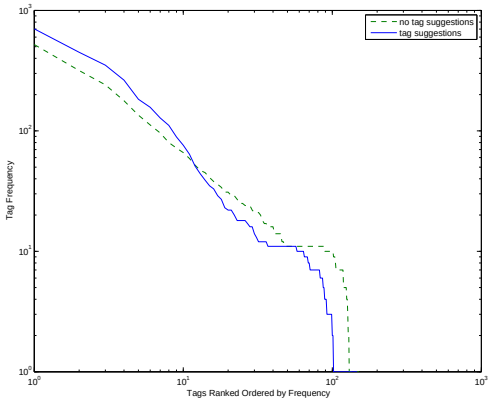

Figure 2. Averaged tag-resource distributions for both experimental conditions on a log-log scale. The solid line depicts the 'tag suggestion' condition, the dotted line the 'no tag suggestion' condition.

1) Parameter Estimation via Maximum-Likelihood: The most widely used method to check whether a distribution follows a power-law is to apply a logarithmic transformation, and then perform linear regression, estimating the slope of the function in logarithmic space to be α. However, this least-square regression method has been shown to produce systematic bias, in particular due to fluctuations of the long tail [12]. To determine a power-law accurately requires minimizing the bias in the value of the scaling exponent and the beginning of the long tail via maximum likelihood estimation. See Newman [13] for the technical details. To determine the α of the observed distributions, we fitted the data using the maximum likelihood method recommended by Newman [13]. Figure 3 shows the different α parameters for the 'tag suggestion' and 'no tag suggestion' conditions, as well as the α determined via aggregation of tagging data from del.icio.us for the 11 URIs. Overall, for the 'no tag suggestion' condition, the average α was 2.1827 (S.D. 0.0799) while for the 'tag suggestion' condition the average α was 2.0682 (S.D. 0.0941). The α values for both conditions and the aggregated data from del.icio.us are situated in the interval $[1.732391 < \alpha < 2.249359]$. Figure 3 shows that both experimental conditions and the aggregated data from del.icio.us have similar exponents. Our results shows that a similar α holds for both the 'tag suggestion' and 'no tag suggestion' condition. Further updates to determine if there is an actual difference between the two conditions as regards if a power-law distribution actually holds.

2) Kolmogorov-Smirnov Complexity: Determining whether a particular distribution is a 'good fit' for a power-

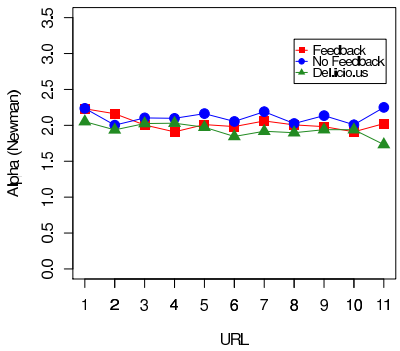

Figure 3. X axis depicts the URI used in the experiment, Y axis depicts the different α values

law is difficult, as most goodness-of-fit tests employ some sort of normal Gaussian assumption that is inappropriate for non-normal power-law distributions. However, the Kolmogorov-Smirnov Test (abbreviated as the 'KS Test') can be employed as a 'goodness-of-fit' test for any distribution without implicit parametric assumptions and is thus ideal for use measuring goodness-of-fit of a given finite distribution to a power-law function. Intuitively, given a reference distribution P (perhaps produced by some well-known function like a power-law) and a sample distribution Q of size n, where one is testing the null hypothesis that Q is drawn from P, then one simply compares the cumulative frequency of both P and Q and then the greatest discrepancy (the D-statistic) between the two distributions is tested against the critical value for n, which varies per function.

For a power-law distribution generating function, we can get a critical p-value by generating artificial data using the scaling exponent α and lower-bound equal to those found in the supposed fitted power-law distribution. A power-law is fit to this artificial data, and then the KS test is then done for each distribution that was artificially generated comparing it to its *own* fitted power-law. The p-value is then just the fraction of the amount of times the D-statistic is larger for the artificially-generated distribution than the D-statistic of the empirically-found distribution. Therefore, the larger the p-value, the more likely a genuine power-law has been found in the empirical data. According to Clauset, "once we have calculated our p-value, we need to make a decision about whether it is *small enough to rule out* the power-law hypothesis" (emphasis added) [12]. The power-law hypothesis is simply that the distribution was generated by a power-law generating function. The null hypothesis is that by chance a function would generate the power-law distribution observed in the empirical data. We shall also use $p \leq 0.1$.

The KS test for all 11 tagged web-pages, testing both the 'tag suggestion' and 'no tag suggestion' condition, is given in Figure 4. The average D statistic for the 'no tag suggestion' condition is 0.0313 (S.D. 0.0118) with $p = .48 (p > .1$, power-law found). For the 'tag suggestion' condition the average D-statistic is 0.0724 (S.D. 0.0256) with $p = .08 (p \leq .1$, no power-law found). These results show that the power-law function exhibited *only* in the 'no tag suggestion' conditions is significant, the fit is closer for the 'no tag suggestion' condition than the 'tag suggestion' condition. The D-statistic showed a range from 0.0170 to 0.0552 for 'no tag suggestion' condition yet a range of 0.0428 to 0.1318 for 'tag suggestion.' Thus, the power-law only significantly appears without tag suggestions, and with tag suggestions a power-law cannot be reliably found. This is surprising, as tag suggestions do not only *not* cause the power-law to form, but they seems that they somehow prevent it from being formed. On the other hand, the 'no tag suggestion' condition results in a significantly good fit to a power-law. Therefore, the result is somewhat counter-intuitive, as according to our experimental data a simple tag-based suggestion mechanism is unlikely the main cause of the power-law formation.

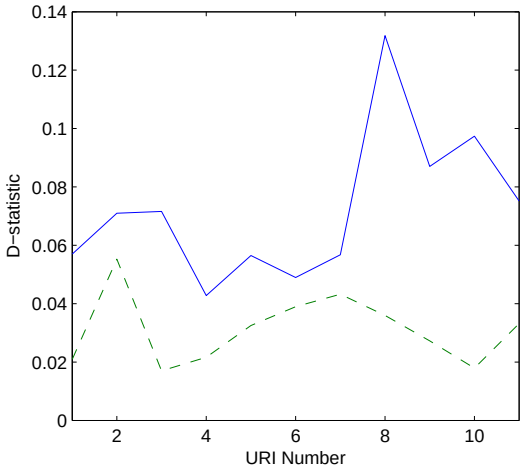

Figure 4. X axis depicts the URI used in the experiment, Y axis depicts the different D Statistics from the KS Test. The dotted line is the 'no tag suggestion' condition, while the solid line is the 'tag suggestion' condition.

B. *Influence of tag suggestion on the tag distribution*

Given that the KS test shows that there is a significant and perhaps counter-intuitive difference in the emergence of the power-law distributions between the conditions, we need a more fine-grained way to tell what the differences are in the distributions for the two conditions. A number of differing techniques will be deployed to answer this question.

1) Kullback Leibler Divergence: The Kullback-Leibler divergence (also known as *relative entropy*), which we abbreviate as 'KL divergence,' can be used an intuitive information-theoretic measure of the distance between two distributions P and Q. Unlike many other methods, it takes

the entire distribution (in our case, the long tail is of particular interest) into account. Note that it is not a true metric as it is an asymmetric, however, it is a useful measure of the difference between two distributions as it is a non-negative, convex function with well-known properties. The KL divergence is zero if and only if the two distributions are the same, otherwise a positive distance results that is larger the greater the divergence between the distributions. Intuitively in information theory, the KL divergence is the expected difference in bits required to encode to distribution Q when using a code based on distribution P. The KL divergence between P and Q is given as:

$$D_{KL}(P||Q) = \sum_{x} P(x) log(\frac{P(x)}{Q(x)}) \quad (3)$$

The KL divergence (using the 'tag suggestion' condition for P and the 'no tag suggestion' condition for Q) for each URI in the experiment are given in Figure 5. While some URIs (like number 6 and 7) have almost no difference between the 'tag suggestion' and 'no tag suggestion' conditions, other URIs like number 11 have large differences. This average KL divergence between the 'tag suggestion' condition and 'no tag suggestion' condition is 0.1617 (S.D. 0.0820). This is small but not insubstantial. As shown in the observation of Figure 2, the long tail of the 'tag suggestion' condition is often shorter than the 'no tag suggestion' condition, while the top of the 'tag suggestion' distribution has a higher frequency than the top of the 'no tag suggestion' distribution. The KL divergence takes this into account, while merely finding the α does not. The effect on the top of the distribution should be investigated further.

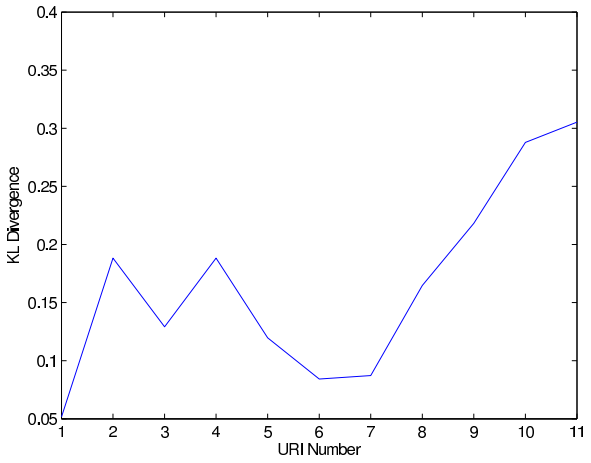

Figure 5. X axis depicts the URI used in the experiment, Y axis depicts the different KL Divergence values

2) Ranked frequency distribution: In order to observe the micro-behavior of the 'tag suggestion' and 'no tag suggestion' distributions, we investigate whether or not the tag suggestion tags are 'forced' higher in the distribution, so leading to a more sparse long tail and an exaggerated top of the distribution in the 'tag suggestion' condition. In order to provide a measurement of the number of suggested tags in the top of the distribution, the percentage of suggested tags that were found in the top 7 and top 10 tags were calculated. We compared the percentage of suggested tags in the top 7 and top 10 ranks for both conditions with del.icio.us. For this we assume that the 7 suggested tags provided by del.icio.us represent the top 7 tags in the ranked frequency distribution so that the percentage of suggested tags in the top 7 and top 10 ranks for del.icio.us is equal to 100%. We averaged the percentages for all URIs per experimental condition.

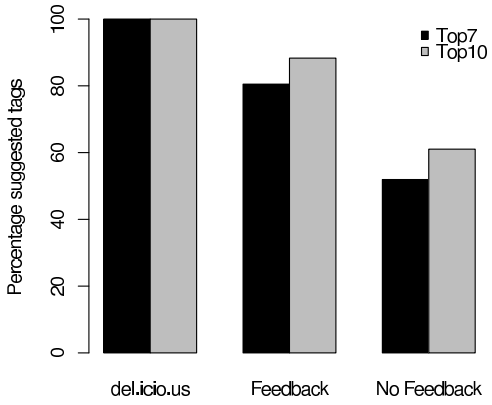

Figure 6. Ranked Frequency Distribution Repeating Suggested Tags

Figure 6 shows that for the percentage of suggested tags available in the top 7 rank for the 'tag suggestion' condition is 80.51% and for the 'no tag' suggestion condition 51.93%. This means that only half of the suggested tags can be found in the top 7 of the ranked frequency distribution in the 'no tag suggestion' condition. So unsurprisingly, in the 'tag suggestion' condition, we observed more of the suggested tags than in the 'no tag suggestion' condition. There is an influence of tag suggestions on the ranked position and the frequency of the suggested tags. Tag suggestions do influence the tag-resource distribution, as tag suggestion causes a net gain of nearly one in three tags being imitated that would otherwise not be. However, when users are not guided by tag suggestions and tag freely they still choose for themselves half of the tags that would have been otherwise suggested had they had a 'tag suggestion' mechanism available. Further we look at the availability of suggested tags in the top 10 as an indication how dispersed the suggested tags are in the ranked frequency distribution for both conditions. For the top 10 rank figure 6 shows that the percentage of suggested tags in the 'tag suggestion' condition is 88.30% and for the "no tag suggestion" condition is 61.03%.

3) Imitation Rates: Another metric that measures the influence of tag suggestion on the tag distribution is the matching and imitation rate as proposed by Suchanek et al. [11]. The matching rate measure the proportion of applied tags that are available in the suggested tags. This metric provides insight in how the user is influenced by the tag suggestion provided by the tagging system. For our experiment the *matching rate* (mr) is being defined as:

$$mr(X) = \frac{\sum_{i=1}^{n} \mid T(X,i) \cap S(X) \mid}{\sum_{i=1}^{n} \mid T(X,i) \mid} \quad (4)$$

X denotes the tag suggestion method that is being used in both our conditions. The 'tag suggestion' condition provides 7 suggested tags while the 'no tag suggestion' condition provided no suggested tags. For a given URI, $T(X,i)$ denotes the set of tags at the ith tag entry and $S(X)$ denotes the suggested tags for that URI. For a tagging instance in which all tags are given by the suggested tags the matching rate will be 1.

The matching rate for the 11 URIs in the experiment and over the both conditions was calculated. The resulting matching rates can be found in Table I. The 'no tag suggestion' condition serves as a reference point. The results in Table I show that users in the 'tag suggestion' condition are being influenced by the appearance of tag suggestions. The average matching rate for the 'tag suggestion' condition is 0.57 (S.D. 0.086) and for the 'no tag suggestion' condition 0.35 (S.D. 0.068). The main drawback of the matching rate is that it can't account for the application of suggested tags when tag suggestion is absent.

Table I
MATCHING RATE

URI No.	Tag Suggestion	No Tag Suggestion
1	0.47	0.31
2	0.57	0.34
3	0.53	0.32
4	0.65	0.48
5	0.45	0.29
6	0.52	0.29
7	0.58	0.38
8	0.65	0.38
9	0.74	0.46
10	0.63	0.30
11	0.59	0.31

This ability to account for tag repetition even when the tag is missing is given by the *imitation rate* (ir), defined as [11]:

$$\alpha_n(S) = \frac{prec_n(X,S) - prec_n(NONE,S)}{1 - prec_n(NONE,S)} \quad (5)$$

With $prec_n(X,S)$ defined as:

$$prec_n(X,S) = \frac{\sum_{i=1}^{n} \mid T(X,i) \cap S \mid [S(X,i) = S]}{\sum_{i=1}^{n} \mid T(X,i) \mid [S(X,i) = S]} \quad (6)$$

The term $prec_n(X,S)$ defines the proportion of applied tags that are available in the single tag suggestion set S. Since the tags S in our experiment is always static, $prec_n(X,S)$ is equal to the calculation of the matching rate for the tag suggestion condition in Equation 4. $prec_n(NONE,S)$ defines the proportion of suggested tags that are available in the tags applied by the user when no tag suggestion is given. This is similar to the calculation of the matching rate for the 'no tag suggestion' condition. Therefore we can rewrite the imitation rate as:

$$ir = \frac{mr(ConditionA) - mr(ConditionB)}{1 - mr(ConditionB)} \quad (7)$$

Table II shows the imitation rates for the different experimental URIs. An imitation rate of 1 will denote full imitation. The results show that users tend to select suggested tags when the are available with a chance of 1 out of 3 with a mean imitation rate of 0.36 (S.D. 0.097).

Table II
IMITATION RATE

URI No.	Imitation Rate
1	0.22
2	0.35
3	0.29
4	0.35
5	0.20
6	0.34
7	0.31
8	0.42
9	0.50
10	0.48
11	0.43

Combining this insight with our previous work in KL divergence and looking at Figure 2, it appears that 'tag suggestion' condition 'compresses' the distribution that naturally arises without tag suggestions. This 'compression' of the distribution that the 'no tag suggestion' generates can be defined as highly frequent tags being reinforced more and less frequent tags reinforced less or not used at all, leading to more imitation in the top of the distribution and a 'shorter' long tail. It is because of this 'compression' caused by tag suggestions that the averaged 'tag suggestion' distributions does not significantly fit power-law distributions while the averaged 'tag suggestion' distribution does fit a power-law distribution. Taking a 'scale-free' power-law as an ideal stable tag distribution, rather counter-intuitively a simple tag suggestion scheme based on frequency may actually hurt rather than help the stabilization of tagging as a power-law distribution.

V. CONCLUSION

The research presented in this paper provides a first step that leads to a new interpretation of the accepted theories and models that explain the emergence of power-laws in tagging systems. Common wisdom in tagging suggested that the

power-law was unlikely to form without tag suggestions. As put by Marlow, Boyd, and others, "a convergent folksonomy is likely to be generated when tagging is not blind," blind tagging being tagging without tag suggestions [14]. The results show that the tags of users *without* tag suggestions converge into a power-law distribution. Moreover, a power-law function fits *more closely* the behavior of users when the users are *not* given tag suggestions than when the users are given tag suggestions. This means that tag suggestions distorts the power-law function that would already naturally occur when users tag blindly without tag suggestions. These results are not unexpected. After all, *words in natural language naturally follow a power-law*, and there exists purely information-theoretic arguments why this is the case [15].

This helps clarify a number of experimental results from previous experiments in tagging. First, this result clarifies how the power-law distribution was observed by Cattuto et al. even before del.icio.us began using tag suggestion via the tag interface [7]. Second, it also helps explain how the majority of users in Suchanek et al.'s experiment had a high matching rate, even when in their report-back most of them said they didn't use or even notice tag suggestions [11]. Our experiment does have a number of limitations, in particular our experiment should be extended to deal with more web-pages as well as expert and non-expert users dealing with different kinds of expert subject matters. In this situation, tag suggestions may have more of an influence on tagging behavior. Although the presented results indicate that some of the previous assumptions underlying the emergence of power-laws do not hold, a power-law distribution alone does not provide the necessary information needed to determine the role of tag suggestion on tag behavior. One line of research that seems promising is to understand how human categorize in general, which could easily influence how they decide which tags to use to annotate web-pages. While the large amount of tagging data on the web made it easy to develop simple mathematical models of human behavior, it seems that a more detailed understanding of what users are *actually* doing is needed.

ACKNOWLEDGMENTS

Dirk Bollen performed his research on the IBBT Project at CUO (Centre for User Experience Research) at the University of Leuven (KUL) and further elaborated on this study in the MyMedia project. Thanks to Aaron Clauset and Cosmo Shalizi for their code used in testing and fitting power-laws on our data.

REFERENCES

[1] T. Berners-Lee, "Universal Resource Identifiers: Axioms of Web Architecture," 1996, informal Draft. http://www.w3.org/DesignIssues/Axioms.html.

[2] J. Hendler and J. Golbeck, "Metcalfe's law, Web 2.0, and the Semantic Web," *Web Semantics*, vol. 6, no. 1, pp. 14–20, 2008.

[3] T. O'Reilly, "What is Web 2.0: Design patterns and business models for the next generation of software," *Communications and Strategies*, no. 1, p. 17, 2007.

[4] S. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006.

[5] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proc. of the 16th Int. World Wide Web Conference (WWW'07)*, 2007, pp. 211–220.

[6] K. Dellschaft and S. Staab, "An epistemic dynamic model for tagging systems," in *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia (HYPERTEXT'08*. ACM Press, 2008, pp. 71–80.

[7] C. Cattuto, V. Loreto, and L. Pietronero, "Semiotic dynamics and collaborative tagging," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 104, no. 5, pp. 1461–1464, 2007.

[8] G. Yule, "A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s." *Philosophical Transactions of the Royal Society of London, Ser. B.*, vol. 213, pp. 21–87, 1925.

[9] H. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, 1955.

[10] S. Bornholdt and H. Ebel, "World Wide Web scaling exponent from Simon's 1955 model," *Physical Review E.*, vol. 64, no. 3, pp. (R)–1 035 104–4, 2001.

[11] F. M. Suchanek, M. Vojnovic, and D. Gunawardena, "Social Tags: Meaning and Suggestions," in *17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, 2008.

[12] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," 2007, http://arxiv.org/abs/0706.1062v1 (Last accessed October 13th 2008).

[13] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, 2005.

[14] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Position paper, tagging, taxonomy, flickr, article, toread," in *Collaborative Web Tagging Workshop at WWW'06*, 2006.

[15] B. Mandelbrot, "An informational theory of the statistical structure of languages," in *Communication Theory*, W. Jackson, Ed. New York, USA: Academic Press, 1953.