Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D4.1
Theoretical tools for modeling and analyzing Collaborative Social Tagging Systems — a stream view —

Period covered: from 01/06/2006 to 31/05/2007
Start date of project: June 1st, 2006
Due date of deliverable: May 31st, 2007
Distribution: Public

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: “Sapienza” Università di Roma
Lead contractor for this deliverable: “Sapienza” Università di Roma

Executive Summary

The objective of this deliverable is to provide a statistical and theoretical analysis of a folksonomy, viewed as a time-ordered stream of metadata associated with a growing number of resources through the tagging activity of an online community of users. Although the stream view is probably the simplest possible projection of the complex tripartite structure of a folksonomy, it allows for interesting statistical analysis and modeling. The methods of complex systems science can be deployed following the usual road-map:

(i) a statistical analysis reveals the emergence of general features, shared by different systems;

(ii) simple minimal models are then proposed in order to capture the key mechanisms acting at the microscopic level, with the goal of reproducing the observed features, by means of numerical simulations as well as, when possible, an analytical approach;

(iii) the comparison of model predictions and experimental data suggests further, more refined measures and/or inspires control strategies, aimed at improving the systems.

The results of such a roadmap, i.e. experimental measures, development of models and of control strategies, are the main subject of WP4. We use the raw data from existing systems, as delivered by WP1, the data from the the applications developed in WP2, as well as the measures and tools developed in WP3.

In this sense, WP4 is the most high-level and comprehensive WP of the TAGora project, since it depends and profits on the results delivered by the other WPs. As a consequence, after the first year of project time, WP4 is not expected to deliver its full potential and achieve conclusive results. However, due to the successful activity of data collection and statistical measurement, the study of folksonomies has been developed quite intensively and some directions for control are already in view. For the sake of clarity, we restrict this report to the stream view of a folksonomy, since this has been the subject of most of our theoretical activity so far. Measurements, analysis and some preliminary control strategies dealing with the graph structure of folksonomies has been moved to the report on D3.1, even if the results presented there could be partially regarded as results of WP4.

Outline of the document

The present report begins with a short review of selected measurements and models proposed in the literature for the analysis of text streams. Such studies have been developed mainly in the context of computational linguistics, but they might in principle be applied also to tag streams. Analogies and differences have been stressed throughout the review, and new models have been proposed to better capture some of the specific features of folksonomy streams.

After this introductory theoretical part, the report focuses on two preliminary experiments with a more applied perspective. On one side, a detailed study of tag “categories” represents a first step to bridge the gap between folksonomies and more structured semantic knowledge management

system. On the other side, a first experiment on folksonomy-aided recommendation strategies has been attempted in order to investigate the possible benefits that could arise from the integration of folksonomies with existing commercial applications.
More in details:

- Chapter 1 contains the definition of tag streams, as well as the most interesting statistical measures performed;
- Chapter 2 contains a review of models proposed in the literature (mostly in computational linguistics) to describe and reproduce the observed statistical measures of text streams. Then, it describes in detail a stochastic process with memory that was introduced by us to model the frequency distribution of tags.
- Chapter 3 describe an experiment of categorization of tags and a corresponding study.
- Chapter 4 propose a first attempt of a recommender system that makes use of the metadata contributed by a virtual social community in a popular web site.

Dissemination of the Results

Parts of the results presented in this deliverable have been published as follows:

- Part of chapter 1 was reported in (Cattuto et al., 2007a). The subsection dedicated to the study of correlations partially appeared in (Cattuto et al., 2006) and (Cattuto et al., 2007b).
- The model with memory, a variation of the Yule-Simon model described in Chapter 2, was presented in (Cattuto et al., 2007b) and (Cattuto et al., 2006).
- Chapter 4 is mainly extracted from (Szomszor et al., 2007).

Contents

List of Figures

Chapter 1

Statistical Analysis of Streams

Folksonomies have been known to exhibit striking statistical regularities and activity patterns. In particular, folksonomies are dynamical systems and each time a user tags a resource, the folksonomy grows: the whole tri-partite network representing the folksonomy is an evolving graph with a complex dynamics.

In order to analyze the dynamical properties of the system, the first and most simple approach is to consider the stream view of folksonomy. In this case, the network structure is disregarded, or better, the network is projected in a zero dimensional space. In the following we shall define more precisely this procedure, which will be referred as the stream view of folksonomy, as opposed to the network view, considered in **D3.1**.

In this section we shall introduce the methods of analysis of macroscopic quantities associated to streams. These quantities are pretty simple to define, nevertheless some of them (eg. the dictionary growth) are hard to be explained. In the following, we shall give an overview of the main measures we have been performing on data streams, relying on their temporal order. Such measurements allow to analyze the development of specific aspects of a folksonomy over time.

1.1 Representing Folksonomies as Streams

In the stream view of a folksonomy, the temporal ordering of posts and tag assignments in collaborative tagging systems is taken explicitly into account (see 1). The stream view can be used to analyze the evolution of specific aspects of a folksonomy over time (e.g. the set of tags). We define the stream representation $\mathbb{F}^S$ of a folksonomy $\mathbb{F}$ as follows:

Definition 1 *The stream representation of a folksonomy $\mathbb{F}$ is a tuple $\mathbb{F}^S := (U, T, R, Y, pt)$ where*

- *pt is a function $pt : Y \to n$ which assigns to each tag assignment (TAS) of Y a temporal marker $n \in \mathbb{N}$. The temporal marker allows an ordering of the TAS data along a time axis.*

Due to the tri-partite character of the TAS (Tag ASsignement) data, different types of streams can be defined: stream of tags, stream of users or eventually stream of resources. Furthermore, a stream might be restricted to a subset of the whole stream by picking up a selection of tags (or resources, or users) and considering only the TAS containing those selected tags (or resources, or users).

The possibility of building streams is based on the fact that most collaborative tagging systems record the physical time of creation of new posts, and make this timestamp available for retrieval. Depending on the system, the timestamp may or may not be updated when a post is modified by the user (re-tagging). On building streams, the timestamp of each post is used to establish post ordering and determine the temporal evolution of the system. Care must be taken when building

streams of TAS: the available timestamp is associated with the post as a unity, so one can safely assume that the timestamp of a TAS is the timestamp of the post it belongs to. However, no temporal ordering of TAS is possible inside a post, so that on building a stream of TAS the local ordering (at the post level) is arbitrary. This is not an issue, because the number of TAS in a post is exponentially distributed and small (Cattuto et al., 2007a), so strict temporal ordering is lacking over a span of a few TAS, only.

To convert a time-ordered sequence of posts into a stream of TAS we map each post of the form `(user, resource, {tag1, tag2, ... })` into adjacent TAS of the form `(tag1, user, resource), (tag2, user, resource), ...` , one for each tag in the post.

Of course, relying on post timestamps yields a reconstruction of the history of the system which is only as much accurate as it is true that posts are left unchanged after having been entered into the system. There is usually no way of detecting and accounting for removed and/or updated posts. It nevertheless rather safe to assume that users behave in a "lazy" way and don't modify posts after creating them for the first time. To date, it is assumed in the literature that post removal or updating have a negligible contribution on the overall evolution of the folksonomy.

As an example, Fig. 1.1 displays the total number of distinct tags N present in the global tag stream of *del.icio.us*, as a function of the stream index n. The data are coming from a large-scale snapshot of *del.icio.us* and the global TAS stream is constructed as described above. The stream index n, playing the role of an "intrinsic" time, is simply the position of a given TAS in the TAS stream. n runs from 1 to the number of total tags assignments, i.e. the sum of the number of tags of all posts (about $1.4 \cdot 10^8$ in this case). For each post added to the system, the "clock" n increases by a number of ticks equal to the number of tags in that post. In terms of the stream index n, a remarkably clean power-law behavior (straight line on a log-log plot) can be observed throughout the full history of the system. This is interesting because the data shown in Fig. 1.1 span a time interval covering almost the entire history of *del.icio.us*: the power-law trend emerges already at the very beginning and is obeyed all the way to present times, as the number of active users and that of bookmarked resources dramatically increase over time. It is worth noticing that the number N of distinct tags does not appear to level off towards a steady-state plateau. This is not surprising in its own merit because tagging systems are open-ended system and new users and resources are a source of continuous novelty for the tags comprised by the folksonomy (Cattuto et al., 2007a).

1.2 Cumulated tag occurrences

An important stream analysis method shows the relative fraction of the cumulated tag occurrences as a function of the age of a resource or user. Usually, the age of a resource or user will be measured in the number of postings assigned to a resource or assigned by a user. The graph shows whether certain tags get more reinforced by users at a resource over the time or whether users develop a certain, stable core vocabulary which might e.g. represent their main topics of interest. In (Cattuto, 2006), the graph was used for showing that the relative proportions of the most popular tags at a resource reach a quite stable state after an initial transient. In (Steels, 2006), it was used for showing the phenomenon that occasionally a new tag may take over already established tags. The latter shows how the vocabulary of a tagging system and of users adapts to important changes in the world or new fashions. Typically, the time is measured in number of postings and not in e.g. days because the popularity of a resource has an important influence on how fast the tag fractions reach a stable state. An example, how such a graph of cumulated tag occurrences can be seen in Fig. 1.2.

For the analysis of the tag fractions for a resource or user, one starts with reducing the overall stream of posts to those related to a single resource or user. After that, one sorts the remaining posts on the basis of their time stamps and accumulates the occurrences of a certain tag over the whole stream.

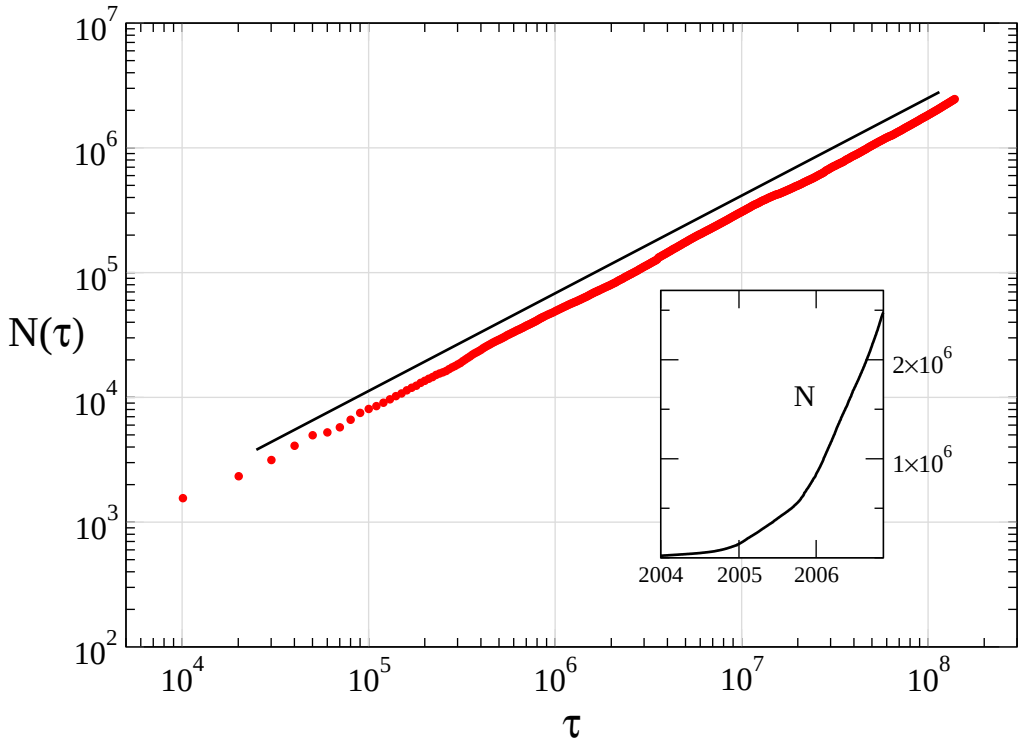

Figure 1.1: Temporal evolution of the total number of distinct tags in *del.icio.us*. As a function of the intrinsic time n (see the definition of tag stream), the number $N(\tau)$ of distinct tags (red dots) increases closely following a power-law (straight line in a log-log plot) across the entire history of the system. The solid black line, provided as an aid for the eye, corresponds to a power-law with exponent $\simeq 0.8$. The inset shows the number N of distinct tags as function of the physical time recorded in post timestamps, spanning almost 3 years of growth and six orders of magnitude in vocabulary size. The main graph and the inset refer to the same interval of physical time.

In chapter 3, the cumulated occurrences of tags will be used for analyzing in how far the emergence of patterns in the tagging behavior of the users can be observed. One of the advantages of tagging systems over e.g. professional annotations is that no specific rules are given how tags are correctly used and that no predefined set of tags exists. Nevertheless, there would be an increased benefit for the users from their collaboration in such a system if they develop a similar behavior with regard to how compound words like “San Francisco” are handled or whether the singular form of a tag is preferred over its plural. An alignment of the user’s behavior would lead to an increased recall when querying the system for resources.

A cumulated occurrences of tags will be also used to characterize the user profile in a folksonomy aided recommendation strategy, whose detailed description will be presented in Chapter 4.

More formally, given a tag stream, the cumulated occurrence of tag, or more simply, the tag frequency, is simply:

$$f_{\text{tag}}(T) = \sum_{t=0,T} \delta(\text{tag}(t), \text{tag})$$

where $\text{tag}(t)$ is the tag at the posistion t of the stream, and $\delta(\text{tag}_1, \text{tag}_2)$ is the Kronecker delta function , taking the value 1 when the two tags are equal and zero otherwise. The relative tag

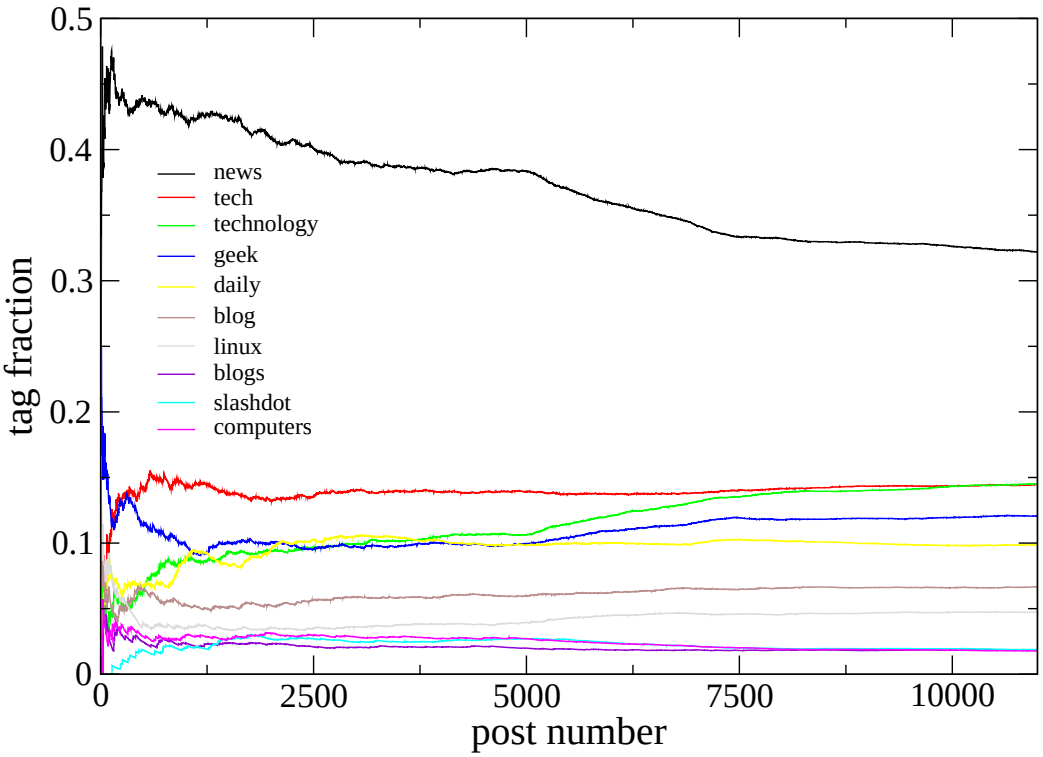

Figure 1.2: Relative fraction of cumulated tag occurrences for `http://slashdot.org` in del.icio.us shown as a function of time, measured by the number of posts associated with the resource (figure from (Cattuto, 2006)).

occurrence, also named as tag fraction or normalized tag frequency, is simply $f_{\text{tag}}(T)/T$.
In the following, when T is the total length of the stream, the frequency $f_{\text{tag}}(T)$ will be called the global frequency of the tag in the stream, or more shortly the tag frequency f.

1.3 Marginal distributions

One of the most used global measures performed on data streams is the plot of the frequency-rank distribution of single elements in the stream. Elements are sorted in a descendent way, according to their number of occurrences in the stream. A graph is plotted with the resulting rank of the elements on the horizontal axis and its number of occurrences (frequency) on the vertical axis. In such a graph the most frequent element is found as the point at the upper left part of the graph, in correspondence of a rank equal to one. The most rare element, usually appearing only once in the stream, would then get a rank equal to the number of total different elements in the stream. It is common practice to consider elements with the same number of occurrences as having different rank according to their sequential position in the stream: if element *gnu* and *bison* have occurred only once, the one appeared first in the stream will get the lower rank. If N is the total number of elements in the stream, it is straightforward to observe that $\sum_{r=1}^{N} p(r) = N$. It can be shown that the frequency-rank distribution, usually indicated as $p(r)$ is connected with the frequency distribution $p(f)$, defined as the measured probability of occurrence of elements with frequency f in the stream. The frequency-rank distribution plot is often referred to as Zipf's plot in honor of G. K. Zipf, who first analyzed in a systematic way the frequency of elements associated to observables referring to human activities (Zipf, 1949). In his analysis, he discovered a remarkable power-law $p(r) \approx r^{-1}$ regularity, independent of the kind of observable he focused in (eg. cities ordered by inhabitants, or words in texts ordered by frequency). Today we know that the

Figure 1.3: *Tag clouds* are a visual device commonly used to show the most-frequently occurring tags in a tagging system. For each tag, the size of the font face is proportional to the logarithm of the frequency of occurrence of the tag within the system. Color encodes the same information, with red for high-frequency tags and blue for low-frequency tags.

strict $1/r$ hyperbolic power-law observed by Zipf is an exception rather then a rule. Often, in fact, the observed curve is a power-law with an exponent differing from -1 (in this case the behavior is referred to as *generalized* Zipf distribution), and the power-law behavior is restricted to limited regions on the rank axis (Mandelbrot, 1959). In order to connect the frequecy-rank distribution $p(r)$ and the frequency distribution $p(f)$ we note that if $p(r) \simeq r^{-\alpha}$ then $p(f) \simeq f^{-\beta}$ with $\beta = 1 + 1/\alpha$. As a result, the Zipf's law $p(r) \simeq r^{-1}$ translates as $p(f) \simeq f^{-2}$ in terms of frequency distribution. The inverse square distribution of frequencies is often named after V. Pareto.

A particularly effective way to display a set of elements according to their frequency of occurrence in a stream is given by tag clouds. Tag clouds show the most frequent elements of the set under consideration by placing them in a frame and printing them choosing a character font face with size proportional to the logarithm of their occurrence. The logarithm is necessary because of the characteristic fat tailed distribution of frequencies. In addition a color code ranging from red to blue, standing respectively for most frequent and less frequent, might be used. A pretty example of tag cloud is presented in Fig. 1.3.

The analysis of streams in terms of frequency-rank distributions might be performed by selecting particular elements and analyze the stream of the elements with a given relation with the selected ones. In this particular case one speaks of *marginal* frequency-rank distributions, i.e. distributions subject to constraint limitations.

As an example taken from folksonomies —with tags, users and resources as stream elements— , we analyzed data from *del.icio.us*and *Connotea* and investigated the statistical properties of tag association. Specifically, we selected a semantic context by extracting the resources associated with a given tag X and study the statistical distribution of tags co-occurring with X (see Table 1.1). Fig. 1.4 shows the marginal frequency-rank distributions for the tags co-occurring with a few selected ones. The high-rank tail of the experimental curves displays a power-law behavior, signature of an emergent hierarchical structure, corresponding to a generalized Zipf's law (Zipf, 1949) with an exponent between 1 and 2. Since power laws are the standard signature of self-organization and of human activity (Barabasi, 2005b; Newman, 2005), the presence of a power-law tail is not surprising. The observed value of the exponent, however, deserves further investigation because the mechanisms usually invoked to explain Zipf's law and its generalizations (i Cancho

Table 1.1: Statistics of the datasets used for the co-occurrence analysis. For each tag in the first column we report the number of posts marked with that tag, the number of total and distinct tags co-occurring with it, and the corresponding number of resources. The data were retrieved during May 2005.

Tag	No. posts	No. tags	No. distinct tags	No. resources
Blog	37974	124171	10617	16990
Ajax	33140	108181	4141	2995
Xml	24249	108013	6035	7364
H5N1	981	5185	241	969

and D.P.Servedio, 2005) don't look very realistic for the case at hand, and a mechanism grounded on experimental data should be sought.

Moreover, the low-rank part of the frequency-rank curves exhibits a flattening typically not observed in systems strictly obeying Zipf's law. Several aspects of the underlying complex dynamics may be responsible for this feature: on the one hand this behavior points to the existence of semantically equivalent and possibly competing high-frequency tags (e.g. *blog* and *blogs*). More importantly, this flattening behavior may be ascribed to an underlying hierarchical organization of tags co-occurring with the one we single out: more general tags (semantically speaking) will tend to co-occur with a larger number of other tags. In this scenario, we expect a shallower behavior for tags co-occurring with generic tags (e.g. *blog*) and a steeper behavior for semantically narrow tags (e.g. *ajax*. To better probe the validity of this interpretation, we investigate the co-occurrence relationship that links high-rank tags, lying well within the power-law tail, with low-rank tags located in the shallow part of the distribution. Our observations point in the direction of a non-trivial hierarchical organization emerging out of the collective tagging activity, with each low-rank tag leading its own hierarchy of semantically related higher-rank tags, and all such hierarchies merging into the overall power-law tail.

As can be seen from the previous example, the analysis of the simple statistical indicators represented by the marginal frequency-rank distributions, may already lead to the understanding of important features of the system. We point out that marginal distributions probe only pure frequency effects of elements in the stream. In particular, they do not provide the possibility to study possible correlations among elements in the stream. In fact, marginal distributions remain unaltered after reshuffling of the elements in the stream. In order to study correlations between element occurrences one must rely on more sophisticate statistical indicators.

Some authors proposed a classification of words in language, based on their rank and frequency. For instance Balasubrahmanyan and Naranan (Balasubrahmanyan, 1996) identify a specific value of the frequency f_0 dividing the vocabulary in two disjoint set of words: for $f < f_0$ many different words appear with a given frequency, i.e. $p(f) > 1$, and the corresponding words are named C-words; for $f > f_0$, on the contrary, only one word appear with the chosen frequency, if any: in the region $f > f_0$ the $p(f)$ is an intermittent discrete function taking values 1 or 0, and the corresponding words are named S-words. The idea is to distinguish between content (C-)words, which constitute the majority of the dictionary but appear few times in the text, and service (S-)words, which are few very frequent words, mainly with gramamtical function. In the case of tag streams this kind of distinction would also be possible, on a purely statistical mathematical sense, but its meaning should be completely different, since tags are all semantic in nature (there is no grammatical structure in tag streams).

More refined statistical study of Zipf's law in texts reveals a richer phenomenology. Mandelbrot, for instance, propose (Mandelbrot, 1953) a slightly modified version row the rank-frequency distribu-

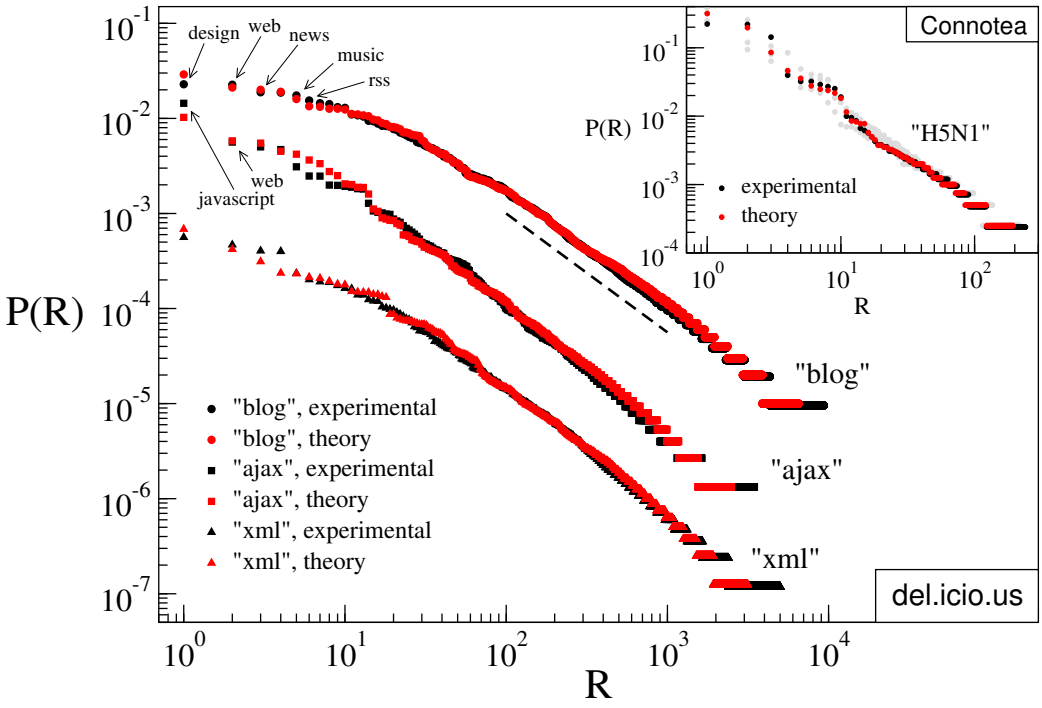

Figure 1.4: Frequency-rank plots for tags co-occurring with a selected tag: experimental data (black symbols) are shown for *del.icio.us* (circles for tags co-occurring with the popular tag *blog*, squares for *ajax* and triangles for *xml*) and *Connotea* (inset, black circles for the *H5N1* tag). For the sake of clarity, the curves for *ajax* and *xml* are shifted down by one and two decades, respectively. Details about the experimental datasets are reported in Table 1.1. All curves exhibit a power-law decay for high ranks (a dashed line corresponding to the power law $R^{-5/4}$ is provided as an aid for eye) and a shallower behavior for low ranks. Some of the highest-frequency tags co-occurring with *blog* and *ajax* are explicitly indicated with arrows. Red symbols are theoretical data obtained by computer simulation of the stochastic process (Cattuto et al., 2007b) described in Section 2.4. Gray circles correspond to different realizations of the simulated dynamics.

tion

$$f(r) = \frac{A}{(r+c)^{\alpha}}$$

which describe a deviation (flattening) of the power law for low-rank words. Similarly Balasubrahmanyan and Naranan propose a modified frequency distribution

$$p(f) = Be^{-\mu/f} f^{-\beta}$$

which deviates from the power law at high frequency. The same authors propose for $p(r)$ a Cumulative Modified Power Law

$$p(r) = \sum_{i=r}^{r_{max}} De^{-\nu/i} i^{-\delta}$$

that they motivate with an information theory model (Balasubrahmanyan and Naranan, 2002). For very large corpora a double slope power law decay is also observed, for instance see Fig. 1.5 taken from (Montemurro and Zanette, 2002).

Although these refined statistical analysis show a quite rich phenomenology, the quest for a simple explanation of the main statistical features observed has been the matter of intense investigation and debates in the last decades (see for instance (zip) for a constantly updated bibliography).

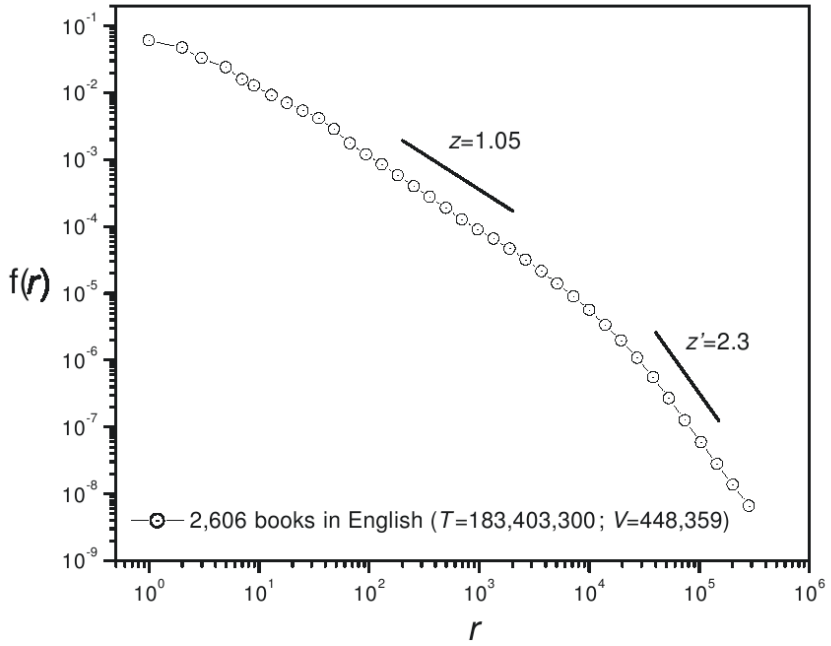

Figure 1.5: Zipf's plot for a large corpus comprising 2,606 books in English, mostly literary works and some essays. The straight lines in the logarithmic graph show pure power laws as a visual aid. Figure and caption from (Montemurro and Zanette, 2002)

Many models have been proposed to explain Zipf's law in language. They range from *monkey random typewriting* models (Miller, 1957), where a power law is obtained without the need of any specific linguistic ingredients, to recent models (i Cancho and Solé, 2003) that try to ground mathematically the original Zipf idea of a principle of least effort for speaker and hearer during a communication.
Two main different approach can be recognized:

- Model based on optimization principles
- Model based on stochastic dynamics

Sometimes the two approach coincide, since the asymptotic statistics for a stochastic dynamics can be rephrased in term of maximization/optimization of a suitable entropy function. However, stochastic models can provide the description of dynamical quantities others than the mere frequency distribution of words.
In the next section we review some of the theoretical models proposed for text statistics and others explicitly suited to tags streams. In the rest of this section, instead, we discuss other dynamical measures that could characterize statistical properties of streams.

1.4 Vocabulary growth

An example of a dynamical quantity that has been measured in text statistics, and which can be easily measured in general on a stream of data, is the vocabulary size.
The quantity represents the number of different words V after a number t of words.

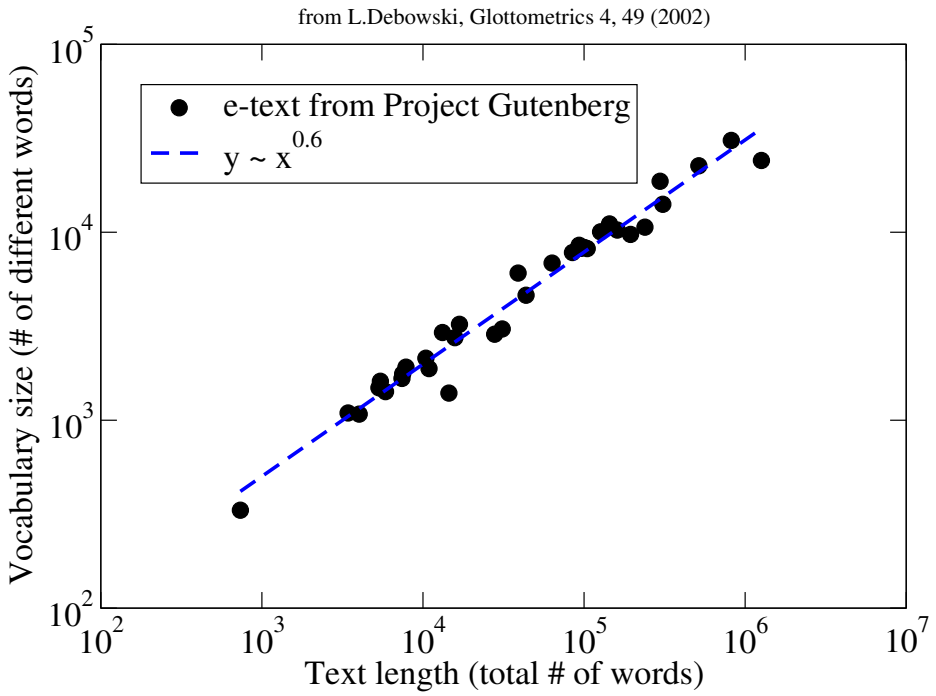

Figure 1.6: Vocabulary size vs. text length in a collection of texts chosen from the English Gutenberg Project

1.4.1 Vocabulary growth in texts

In texts this quantity has been studied, for instance in (Debowski, 2002; Gelbukh and Sidorov, 2001; Kornai, 2002).

Fig. 1.6 shows the analysis of the total number of different words (vocabulary size) as a function of the text length (total number of words) for a collection of 35 e-texts from the English Project Gutenberg (Li, 2006) (data from (Kornai, 2002)). As can be seen, the data show a degree of regularity, and a possible description is that the size of the vocabulary V increase with the length of the text t as

$$V(t) \propto t^{\gamma} \tag{1.1}$$

This law is also known as Heaps law in texts and the exponent γ is usually smaller than one. This behavior seems quite universal: similar studies performed on different corpora show very similar behavior. However the numerical values of the exponent seems to depend on language (Gelbukh and Sidorov, 2001). In Fig. 1.7 we show the results of their measures performed on a sample of English and Russian texts. The authors measured the Zipf exponents on two corpora of 39 texts each. At the same time, for each text, they consider the growth of the number of different wordforms or lemmas n_i as a function of the running word number i, and again they observed a Heaps law in the form:

$$\log n_i \approx D + \gamma \log i$$

The numerical values of the measured exponents show a slight difference between the two languages, as well as a certain dispersion for each language. Interestingly the two exponents show a degree of correlation. In the inset we considered the reciprocal of γ versus α and we observed a kind of linear correlation $\alpha \propto 1/\gamma$.

Interestingly, the exponents measured on English texts are slightly larger than the exponent measured in Fig. 1.6. There is, in fact, a slight difference in the two exponents. The exponent of Fig. 1.6 describe the scaling of the total number of different words as a function of the total length of the text *for several texts*. On the other hand, each point in Fig. 1.7 represent the exponent fitted in

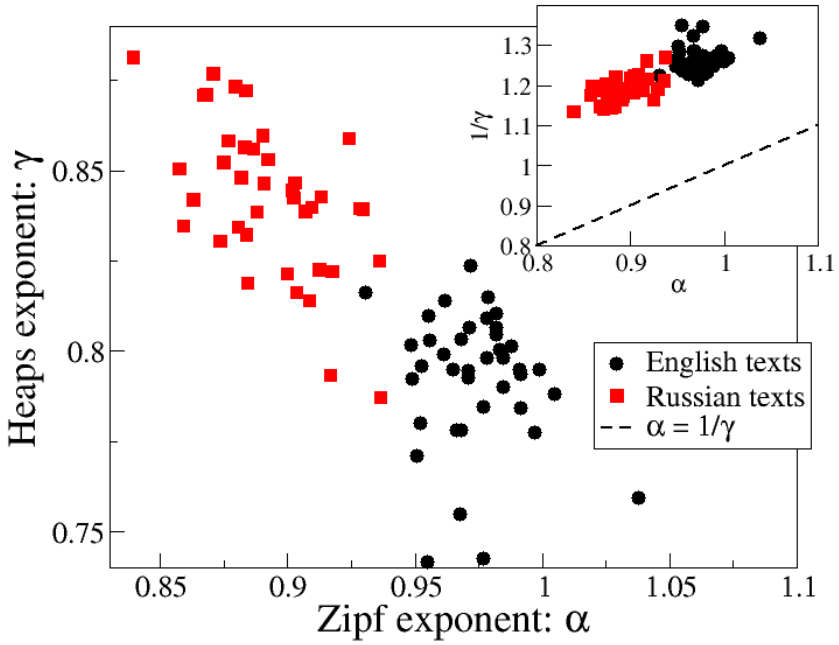

Figure 1.7: Zipf and Heaps exponents for English and Russian texts. Data from (Gelbukh and Sidorov, 2001). In the inset we compare the Zipf exponent with the reciprocal of the Heaps exponent. In the case of an uncorrelated stationary stream the vocabulary should grows as Eq. (1.2), and the corresponding relation should be $\alpha = 1/\gamma$ (dotted line in the inset).

the vocabulary growth *for a single text*. The discrepancy seems to suggest that the scaling (1.1) is not perfect and a change in the slope could happen for very long text, in analogy with the slope observed for rare words in the Zipf law.

The power law growth of the vocabulary is not surprising when the rank frequency function follow a Zipf's law. A simple argument can give an estimation of the exponent γ as function of the exponent α. In fact, for a stationary uncorrelated streams, the typical time of arrival of a tag is the inverse of its probability, which is proportional to the observed frequency:

$$t \propto \frac{1}{f}.$$

This means that at that time one could roughly expect that every word more probable have arrived at least once. Hence the number of different tags appeared in the stream is nothing but the rank r. Since $f \propto r^{-\alpha}$ this gives an expected growth of vocabulary:

$$V(t) \propto t^{1/\alpha}, \tag{1.2}$$

i.e. $\gamma = 1/\alpha$ for an uncorrelated stationary stream satisfying Zipf's law.

However the result is different from the behavior observed in the inset of Fig. 1.7 (the dashed line being the simple prediction (1.2)) witnessing that the sub-linear growth of the vocabulary size is a sensible measure of inner correlation in texts.

1.4.2 Vocabulary growth in folksonomy

A very similar observation has recently be made in folksonomy streams (Cattuto et al., 2006). In this case it is possible to measure how the number of different tags increase with the length of the

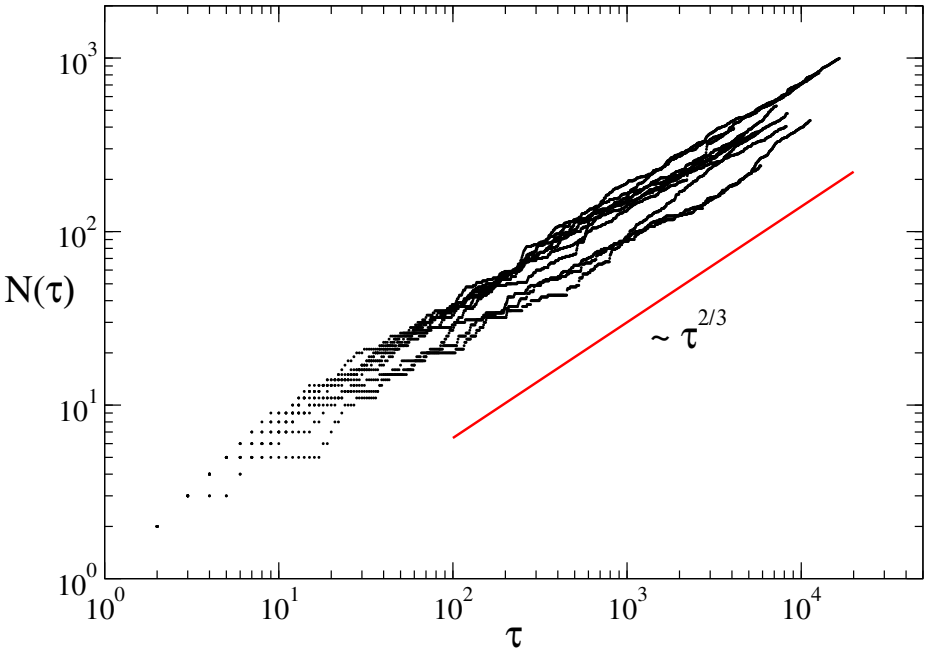

Figure 1.8: Vocabulary growth for single resources. For 10 different resources in *del.icio.us*, the number of distinct tags $N(\tau)$ associated with them is plotted as a function of the intrinsic time τ pertaining to each resource. While single resources display a somehow noisy evolution, an overall power-law behavior governing the vocabulary growth is apparent, with an exponent $\gamma \simeq 2/3$ (red line).

stream. As shown in Fig. 1.1, the total number of distinct tags plays the role of the vocabulary size in texts. An impressive Heaps law is observed with quite large exponent $\gamma > 0.8$.

A more detailed analysis can be carried on different tag streams, using the available `del.icio.us` dataset. To this end, we consider the growth of the number $N(\tau)$ of distinct tags associated with the 10 popular resources, as a function of the intrinsic (resource-specific) time τ. The resources are chosen among the 1000 top-bookmarked resources in the system, starting from rank 100 and decreasing at intervals of 100. While the vocabulary growth exhibits a somehow noisy temporal evolution, the general trend of growth appears to be compatible with an algebraic law of growth, a power-law with an exponent close to $2/3$. This is a striking regularity, valid for very different resources across the system. Also, at this level of detail, no systematic dependence on the popularity of a resource can be detected. The local exponent of growth is smaller than the global one (Fig. 1.1) and the relation between the two may be linked to the statistical properties of tag co-occurrence, and might ultimately provide insights into the semantic structure of folksonomies.

To better probe the similarity of growth behaviors for different resources, we defined a rescaled growth curve, where both the intrinsic time τ and the final number of distinct tags $N(\tau_{max})$ are divided by their final values, τ_{max} and $N(\tau_{max})$, respectively. In this way, the curves for different resource can be easily plotted on the same graph. As shown in Fig. 1.9, all the rescaled curves lie between two limit power-laws, $(\tau/\tau_{max})^1$ and $(\tau/\tau_{max})^{1/2}$. More importantly, all curves tend to lie along a "universal" growth curve with an exponent close to $2/3$.

In order to make a more quantitative measure over a broader set of resources, we implement the following unsupervised procedure for characterizing the growth of local tag vocabularies: for each resource we measure an effective exponent γ that approximates the rescaled vocabulary growth with a power-law $(\tau/\tau_{max})^\gamma$. The simplest way to do this is to compute γ as $\gamma = \log(N(\tau_{max}))/\log(\tau_{max})$. Fig. 1.10 shows the probability distribution of the resulting values of γ, measured for three groups of resources. In particular, the red curve in Fig. 1.10 displays

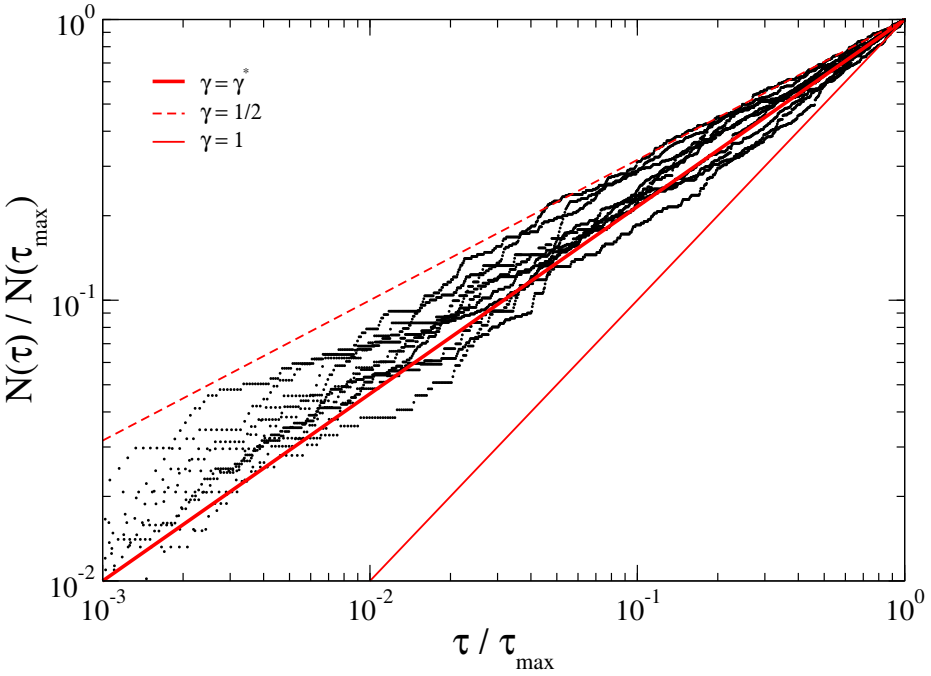

Figure 1.9: Rescaled vocabulary growth. The curves of Fig. 1.8 were rescaled by dividing both the intrinsic time τ and the number of distinct tags $N(\tau)$ by their final (resource-specific) values τ_{max} and $N(\tau_{\text{max}}$, respectively. After rescaling, all curves lie approximately along the "universal" $(\tau/\tau_{\text{max}})^{2/3}$ line (thick red line). On approaching the common endpoint, the slope of all curves appear to lie in the 0.5-1 range (dashed line and thin red line, see also Fig. 1.10).

the distribution of γ values for the 1000 top ranked (most bookmarked) resources in *del.icio.us*. The distribution is well approximated by a rather narrow Gaussian distribution, whose average value is $\gamma^* \simeq 0.7$. This seems to confirm the idea (Fig. 1.9) that there is a well-defined exponent of growth governing the temporal evolution of popular resources. Moreover, the vocabulary growth of popular resources appears slower than the system-wide vocabulary growth of Fig. 1.1.

On computing the distribution $P(\gamma)$ for less and less popular resources (black and blue curves), the distribution gets broader and its peak shifts towards higher values of γ, indicating that the growth behavior is becoming more and more linear. This crossover from sub-linear to linear growth for resources bookmarked by just a few users is expected and corresponds to a sort of "priming" effect for the resource: the first few users who bookmark it build the "core" tag vocabulary for the resource, and since only a few posts are present at that time, most tags are new and the size of the vocabulary grows linearly with the total number of tags τ as well as with the number of posts associated with the resource. As more and more users bookmark the resource, correlations and social effects come into play and the law of growth crosses over from the linear to the "universal" sub-linear behavior reported above.

To make contact between local vocabulary growth in the context of a single resource and vocabulary growth in the context of a single user, we repeat the above analysis for the 1000 most active users in *del.icio.us* (as measured by the number of resources they bookmarked). The resulting probability distribution $P(\gamma)$ is shown in Fig. 1.11 and is qualitatively similar to the ones of Fig. 1.10. In particular, we notice that the peak of $P(\gamma)$ is compatible with the value γ^* observed for the top-ranked resources.

We would like to remark that the huge variability of vocabularies, at the level of single users and resources, is not in contrast with very regular – and simple – features at the global level. On the contrary, the emergence of regularity from the uncoordinated activity of users is the hallmark of complexity and indicates that tools and concepts from complex system science may prove valuable

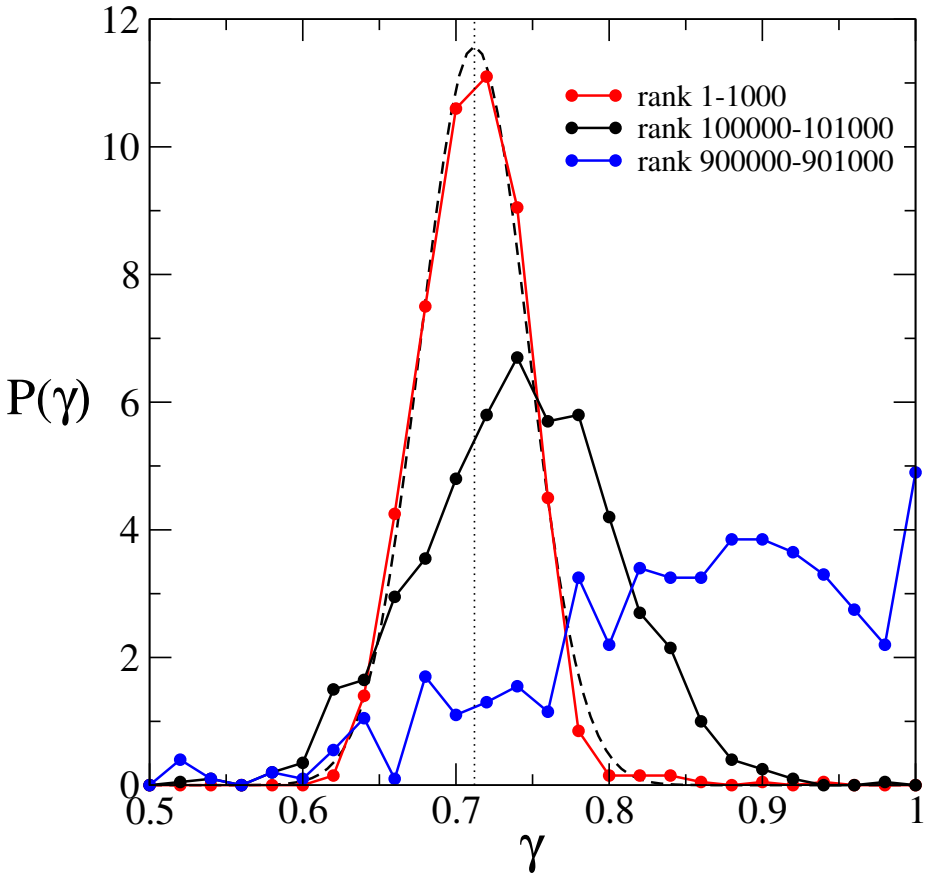

Figure 1.10: Probability distribution of the vocabulary growth exponent γ for resources, as a function of their rank. The red curve is the normalized probability distribution $P(\gamma)$ for the 1000 top-ranked (most bookmarked) resources in *del.icio.us*. It appears to be sharply peaked at a characteristic value $\gamma^* \simeq 0.71$ (vertical line) and can be closely fitted with a Gaussian (dashed line). This indicates that highly bookmarked resources share a characteristic law of growth, as already pointed out in Fig. 1.9. On computing the distribution $P(\gamma)$ for less and less popular resources (black curve and blue curve), the peak shifts towards higher values of γ and the growth behavior becomes more and more linear. The typical number of users who have bookmarked the resources used in this analysis is approximately a few thousands for the red curve, a few hundreds for the black curve, and just a few users for the blue one.

for understanding the structure and dynamics of folksonomies.

These observations point out that sub-linear dictionary growth is a genuine non-trivial feature of the system and open several problems. Is sub-linear growth at the global level (or at the local level) related to correlations among users' activity? Does the growth observed in the context of a single user reflect a collective/cooperative phenomenon, or is it just mirroring the complex cognitive processes (incorporating semantics) at the level of that individual user? Is the difference between local and global exponents relevant, and if so, what kind of information about the structure of tag space is it disclosing? What are the key elements in the user-system interaction that lead to the observed behaviors?

1.5 Correlation functions

Deviations from the trivial scaling (1.2) can be the signature of non-trivial correlation in the stream. For texts, such correlation are obviously expected, due to the internal syntactic structure of lan-

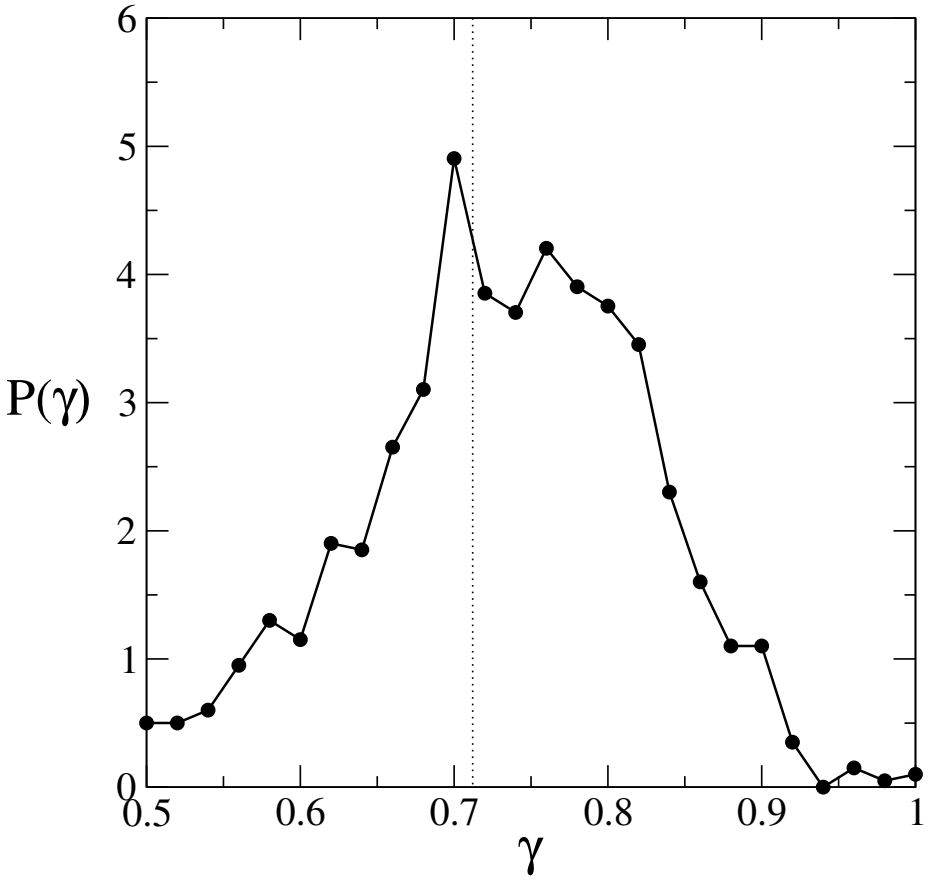

Figure 1.11: Probability distribution of the vocabulary growth exponent γ for user vocabularies. The distribution $P(\gamma)$ was computed for the 1000 most active users in *del.icio.us*. Similarly to Fig. 1.10, it appears peaked around a characteristic value close to the same observed for top-ranked resources (vertical line, same as in Fig. 1.10).

guage.

In a given stream of elements, the order of elements has usually great relevance. A typical extreme example is represented by the stream of words in a text. In fact, the random scrambling of words results generally in a senseless unreadable sequence. This happens because lexical elements of a language have to follow a precise order. In English sentences, for example, articles must precede nouns, subject precedes the predicate, etc. In order for sentences to make sense, this natural developed word order has to be observed: so to say, words must correlate each other. Correlations try to catch the importance of element order in a stream. Written texts in any language represent a limit example of how correlations are important. On the other end, streams extracted from folksonomies, eg. streams of tags, do not have the purpose to transmit any particular meaning to the reader, simply because it is supposed there is nobody reading them. Streams of tags have not the function to be read by humans as they do with fiction. Nevertheless, a certain regularity in element occurrences, apart from mere frequency effects, might be present in the stream.

Montemurro and Pury try to quantify correlations in texts in (Montemurro and Pury, 2002). They associate to each word in the text the corresponding rank and they normalize the resulting time series as to have zero average and unitary standard deviation. In other words they define the quantities

$$\xi(t) = \frac{r(t) - < r >}{\sigma}$$

where $< r >$ is the average rank of the text $< r >= \sum r(t)/T$ and σ the corresponding standard deviation. Then they consider the stochastic process defined as:

$$X(t) = \sum_{u=1}^{t} \xi(u)$$

and measure several two-times quantities, essentially based on the diffusion properties of a walker whose position is $X(t)$. They observe long range fractal correlations, characterized by an Hurst exponent, which slightly depend on the analyzed corpora but always larger than 0.6, to be compared with the exponent corresponding to un uncorrelated time series equal to $1/2$.

The general way to define N-point correlations in a certain finite temporal window from a mathematical point of view is

$$C_{\{A_i\}}(\{\Delta t_i\}, t_0) = \frac{1}{T - \max\{\Delta t_i\}} \sum_{t=t_0+1}^{T-\max\{\Delta t_i\}} \prod_{i=1}^{N} A_i(t + \Delta t_i) \tag{1.3}$$

where t_0 is a time offset, $\Delta t_1 = 0$ and A_i are observables associated to stream elements or simply the elements themselves in case of numerical data streams. Since correlations may change in time, i.e. two strongly correlated quantities may become less tight in future, one can start their analysis at different times t_0. Our study of correlations inside streams focused on the two-point correlation calculated into a finite temporal window of width T

$$C_{A,B}(\Delta t_i, t_0) = \frac{1}{T - \Delta t} \sum_{t=t_0+1}^{T-\Delta t} A(t)B(t + \Delta t) \tag{1.4}$$

We analyzed the 2-point correlations in the case of streams of tags in folksonomies, aiming at understanding how users choose tags in time (Cattuto et al., 2007b). Is there any memory effect that results in newer tags to be preferred to older tags? To answer this question, we moved from the observation that actual users are exposed in principle to all the tags stored in the system (like in the original Yule-Simon model (Simon, 1955)) but the way in which they choose among them, when tagging a new resource, is far from being uniform in time. It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones. Indeed, recent

findings about human activities (Barabasi, 2005b) support the idea that the access pattern to the past of the system should be fat-tailed. Fig. 1.12 shows the temporal auto-correlation function for the sequence of tags co-occurring with *blog*. Such a sequence is constructed by consecutively appending the tags associated with each post, respecting the temporal order of posts. Correlations are computed inside three consecutive windows of length T, starting at different times t_w,

$$C(\Delta t, t_w) = \frac{1}{T - \Delta t} \sum_{t=t_w+1}^{t=t_w+T-\Delta t} \delta(\mathsf{tag}(t + \Delta t), \mathsf{tag}(t)) \,,$$

where $\delta(\mathsf{tag}(t + \Delta t), \mathsf{tag}(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t + \Delta t$. From Fig. 1.12 it is apparent that the correlation function is non-stationary over time. Moreover, for each value of the initial time t_w a power-law behavior is observed: $C(\Delta t, t_w) = a(t_w)/(\Delta t + \gamma(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor and $\gamma(t_w)$ is a phenomenological time scale, slowly increasing with the"age" t_w of the system. $c(t_w)$ is the correlation that one would expect in a random sequence of tags distributed according to the frequency-rank distribution $P_{T,t_w}(R)$ pertaining to the relevant data window. Denoting by $R = R\mathsf{max}(T, t_w)$ the number of distinct tags occurring in the window $[t_w, t_w + T]$, we have $c(t_w) = \sum_{R=1}^{R=R\mathsf{max}(T,t_w)} P^2_{T,t_w}(R)$.

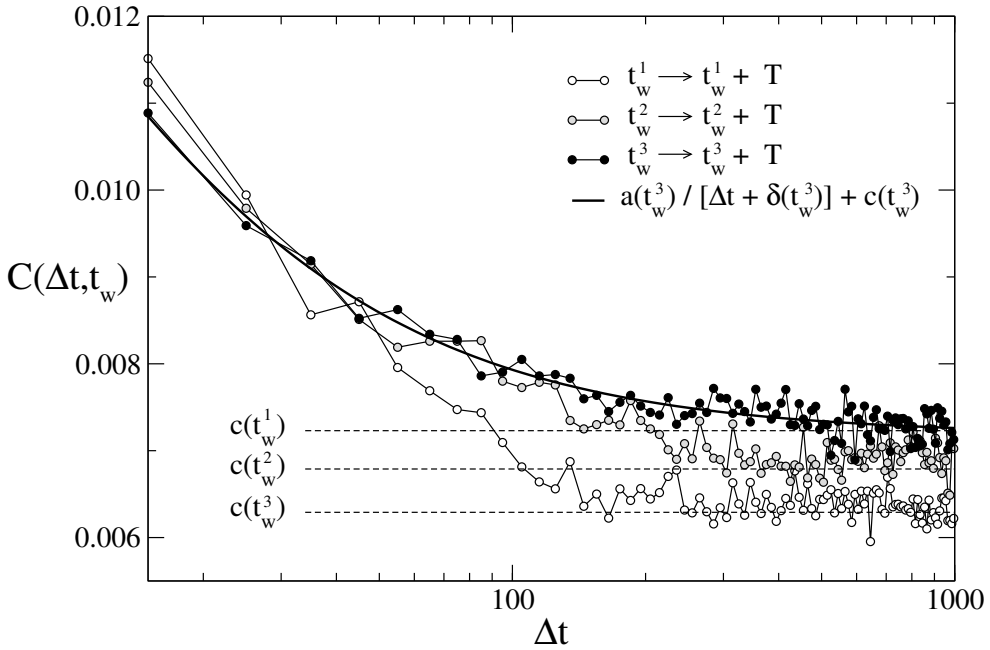

Figure 1.12: Tag-tag correlation functions and non-stationarity. The tag-tag correlation function $C(\Delta t, t_w)$ is computed over three consecutive and equally long ($T = 30000$ tags each) subsets of the *blog* dataset, starting respectively at positions $t^1_w = 10000$, $t^2_w = 40000$ and $t^3_w = 70000$ within the collected sequence. Short-range correlations are clearly visible, slowly decaying towards a long-range plateau value. The non-stationary character of correlations is visible both at short range, where the value of the correlation function decays with t_w, and at long range, where the asymptotic correlation increases with t_w. The long-range correlations (dashed lines) can be estimated as the natural correlation present in a random sequence containing a finite number of tags: on using the appropriate ranked distribution of tag frequencies within each window (see text) the values $c(t^1_w)$, $c(t^2_w)$ and $c(t^3_w)$ can be computed, matching the measured plateau of the correlation functions. The thick line is a fit to the fat-tailed memory kernel described in the section 2.4.

Chapter 2

Minimal Stochastic Models

2.1 Monkey typing model

The minimal stochastic model for tag streams, is the so called Monkey typing model, originally proposed by Miller and Mandelbrot for texts. In this model each character in the stream is thrown with a probability constant in time and independent by the other characters. The "space" character delimits the tags and its probability is p. In the simplest version, the other n characters share equally the rest of the probability $1-p$. The probability of each of the n^l tags by l letters is then:

$$p\left(\frac{1-p}{n}\right)^l.$$

As a consequence, the longer the tag, the smaller its probability. This makes the ranking for decreasing probability coinciding with the rank for increasing length. More precisely, a tag of length l will have a rank r

$$\frac{n^l-1}{n-1} < r < \frac{n^{l+1}-1}{n-1}.$$

For instance, the tag with rank $r = n^l$ has surely length l, as can be easily verified. Its frequency in a long stream is proportional to its probability, and this gives:

$$f(r) \propto p\left(\frac{1-p}{n}\right)^{\log_n r} = p\, r^{-1+\log_n(1-p)},$$

i.e. a Zipf's law with $\alpha = 1 - \log_n(1-p)$.

2.2 Fixed distribution model (FDM)

Although it's true that very frequent words are short tags, it's difficult to ascribe to the monkey typing model a really descriptive value and in fact it has been the subject of some comments about its value in describing real texts statistics. Here we could just note that accordingly to that model the frequent tag "blog" should be equiprobable with "golb" or "glbo", which is not the case.

However the model is interesting as a "zero" or "null" model, in order to understand the relevance of a statistical measure. An other example of zero model is what we called the Fixed Distribution Model (FDM).

The model aims to understand the influence of correlations in tag stream statistics. Here we assume that tags in the stream are chosen independently following a given frequency distribution law. That is, if we call $p(r)$ the probability of the $r-th$ tag, the stream will be a sequence of tags randomly and independently thrown from the distribution $p(r)$ (see Fig. 2.1 for a graphic representation of the model definition).

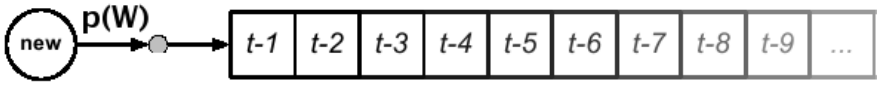

Figure 2.1: Graphic representation of the Fixed Frequency Distribution model. The tags/words composing the stream are thrown independently according to a common time-independent distribution p.

Since, without loss of generality, tags can be sorted for decreasing probability, it's clear that $p(r)$ will be simply proportional to the rank frequency distribution. How will the vocabulary grow during the stream? A very rough argument gives the correct answer. If $V(t)$ is the number of different tags in a stream of length t. Since these tags will be preferentially between the most probable tags, the probability that $V(t)$ increases will be proportional to the probability to throw a less probable tag, i.e. a tag with a rank larger than $V(t)$. In other words (and considering for easy of computation both t and r continuous variables):

$$\frac{dV}{dt} = \int_{V(t)}^{M} p(r)dr \tag{2.1}$$

where M is the total number of different tags one that can ever enter in the stream.

This formula states a little more formally the idea used in the previous section to argue that the power law growth of vocabulary size can be expected in the case of a Zipf's Law. In fact, if $p(r) = Ar^{-\alpha}$, then

$$D(t) = \left(\frac{\alpha A}{1-\alpha}\right)^{\frac{1}{\alpha}} t^{\frac{1}{\alpha}}$$

Interestingly, equation 2.1 gives the correct growth even in the case of M equiprobable tags, that is for $p(r) = 1/M$. In this case, it gives:

$$D(t) = M\left(1 - \exp(-t/M)\right)$$

that for very large M results in a linear growth $D(t) \propto t$. As a opposite case, if $p(r)$ decrease exponentially fast at large r, the vocabulary grows logarithmically in time.

2.3 Yule-Simon model

In order to explicit consider correlated stream of texts, Simon proposed the following stochastic model (Simon, 1955) that can be described as a dynamical construction of the stream. The model depends on a single parameter p, giving the probability to find in a position t of the stream, a word that is not present in any previous position of the stream. That is, At each time t, the new word entering in the stream is, with probability p, a word that was not present in the stream up to that time. In other words, p is the constant rate of production of new words along the stream, i.e. the rate of (linear) growth of the vocabulary size:

$$V(t) = pt \tag{2.2}$$

Otherwise, with probability $1-p$, the word entering at time t is one of the old words. The choice of what old words is made sampling a previous time and copying the tag in the stream at that time. In other words, the previous tag is chosen with a probability proportional to the number of its previous occurrence.

A graphic picture summarizing the dynamical rules of the Yule-Simon model is shown in Fig. 2.2

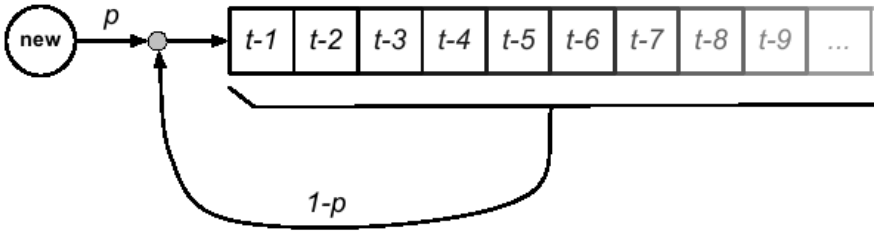

Figure 2.2: Graphic representation of the Yule-Simon model. The only parameter is the constant probability p, which correspond to the rate of linear growth of the vocabulary.

A very sketchy analytic computation of the asymptotic frequency distribution of the stream is the following: Denoting by $n_1(t)$ the number of different words appearing in the stream only once after t, one can write down a dynamical equation:

$$\frac{dn_1}{dt} = p - (1-p)\frac{n_1}{t} \tag{2.3}$$

The number increase with a constant rate p and decrease with a rate proportional to $(1-p)n_1(t)/t$, which correspond to the probability that the word has been chosen as an old word at time t. Similarly, $n_f(t)$, which is the number of different words appearing exactly f times after t words, increases if a word appearing $f-1$ will be chosen at time t and decreases if a word appearing f times will. That is:

$$\frac{dn_f}{dt} = (1-p)\left[\frac{(f-1)n_{f-1}}{t} - \frac{fn_f}{t}\right] \tag{2.4}$$

The simplest solution is obtained considering that, asymptotically, the distribution of words reaches a stationary state, and hence

$$n_f = \frac{p_f}{t}$$

where p_f is the constant fraction of words occurring f times. From equation (2.3) straightforwardly follows that $p_1 = \frac{1}{2-p}$. For $f > 1$, instead, equation (2.4) gives a recursive relation

$$\frac{p_f}{p_{f-1}} = \frac{f-1}{f+1/(1-p)}$$

that admit an exact solution which asymptotically results:

$$p_f \propto f^{-1-1/(1-p)} \tag{2.5}$$

In terms of the rank frequency distribution, this corresponds to a Zipf's law with exponent $\alpha = 1-p$. As shown in Fig. 2.3, although the power law decay is recovered, the exponent is smaller than one, hence smaller than the observed exponents in texts (and tag streams).

The basic mechanism that provides the Zipf statistics is the reinforcement of probability of ancient words along the stream, that is the fact that the more a word enter the stream, the more probable it will enter in the following. This kind of mechanism has been proposed in many different models after Simon, and lately named "preferential attachment" dynamics, in the recent complex network literature. However it can be seen as a restatement of an ancient idea used by Willis and Yule to analyze the data on frequency distribution of the sizes of biological genera.

At odds with the previous presented model, the Yule-Simon model does not assume a frequency distribution from the beginning, neither a fixed distribution for characters nor a distribution for word frequencies. In fact, the probability to observe a word at a certain time t depends on the previous tags, as can be explicitly measured considering the time correlations in the stream. In

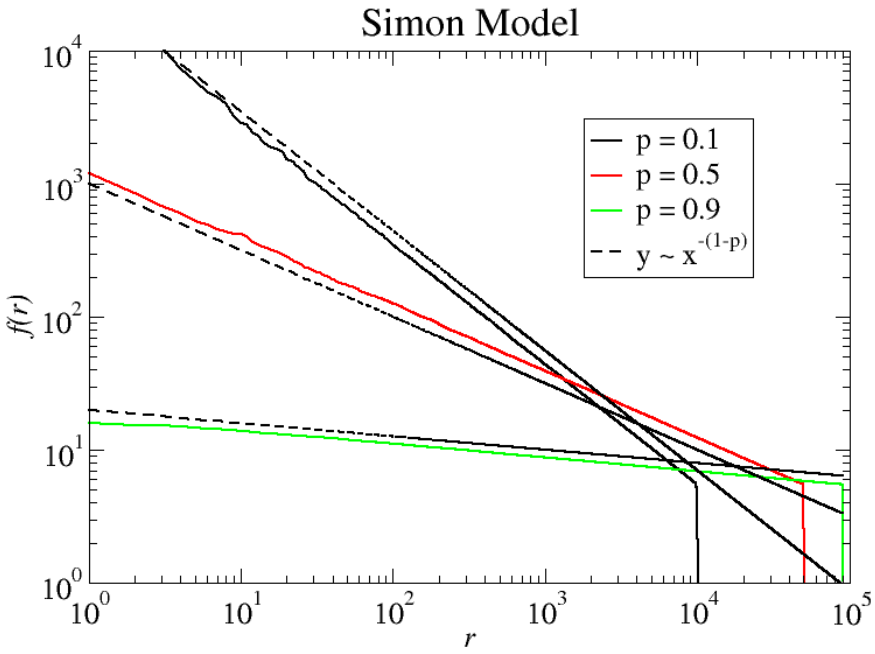

Figure 2.3: Rank frequency distribution for streams generated with Yule-Simon model. The exponent is given by $1 - p$ and it's always smaller then the canonical Zipf's value 1.

Fig. /refsimon-correl is shown the time correlations for two stream corresponding to two different values of the parameter p. As can be noted the correlations keep quite constant for a time of order $1/p$ and then slowly decay to zero. The time $t \simeq 1/p$ represents a cross-over between two different regimes of the dynamics.

For $p = 0$, Yule-Simon model is equivalent to an urn model previously considered by Markov (Markov, 1951) and then studied by Polya (Johnson and Kotz, 1977): an urn containing a number of colored balls is used to produce a stream of events X_1,...,X_n, each one defined as the result of the extraction of a ball from the urn. If after each extraction the extracted ball is replaced in the urn and an other ball of the same color, the stream of events become time correlated, since the probability of each extraction depends on the previous events (for instance $P(X_1, X_2) \neq P(X_1)P(X_2)$, which determines the internal composition of the urn. Interestingly this sequence of events represent an example of an infinite sequence of exchangeable events (Feller, 1968). In fact, it's quite easy to show that different sequences obtained with a permutation of the same set of events, are equiprobable: $P(X_1, X_2, ...X_t) = P(X_{i_1}, ..., X_{i_t})$, where $i_1, ...i_t$ is a generic permutation of the indexes. A stream fulfilling this properties displays constant time correlation, exactly has happens for the Yule-Simon model for $t << 1/p$.

In order to overcome the linear growth of vocabulary and the effective double slope observed for large corpora, Zanette and Montemurro (Zanette and Montemurro, 2005) introduced an *ad-hoc* modified version of the Yule-Simon model, where the rate of growth of the vocabulary is by assumption the correct one:

$$p = p_0 t^{\gamma - 1}.$$

Furthermore, they put a threshold in the probability to get an old world, which increase the probability to choose low frequent words in the stream (this introduce other two parameters). As expected these two modifications makes both the rank frequency distribution and the vocabulary growth generated by the model much closer to the ones observed in texts. In particular they compute the

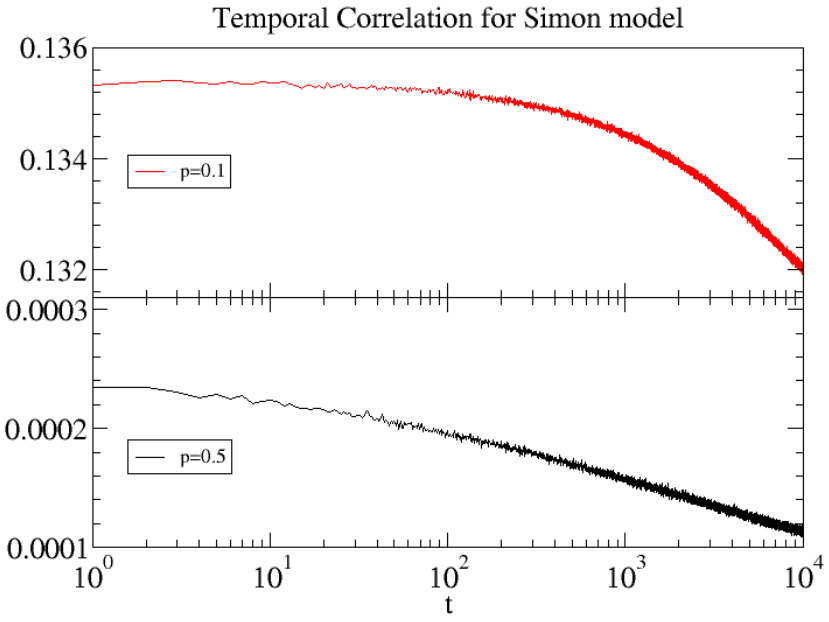

Figure 2.4: Time correlation for two streams generated with Yule-Simon model. The correlation keeps almost constant up to time or order $1/p$, then it decay very slowly.

predicted Zipf's exponent that turns to be the same predicted by the Fixed distribution model:

$$\alpha = 1/\gamma$$

which describe the rank frequency distribution up to a cut-off depending on the parameters of the model.

2.3.1 Preferential attachment measures

In Yule-Simon model, new elements are added to a stream with constant probability p at each time step, whereas with complementary probability $\bar{p} = 1 - p$ an already occurred element is chosen uniformly from within the already formed stream. The same mechanism is at play in the preferential attachment (PA) model for growing networks proposed by Barabási and Albert (Barabási and Albert, 1999). In that case, a network is constructed by progressively adding new nodes and linking them to existing nodes with a probability proportional to their current connectivity. Simon's processes and PA schemes are closely related to each other and a mapping between them has been provided by Bornholdt and Ebel (Bornholdt and Ebel, 2001).

It can be important to understand whether this simple weak correlation rule of the type *rich-gets-richer* is taking place (and to which extent) in the streams we analyze. In order to check for deviations from PA, it is possible to adopt an elegant and efficient way suggested by Newman (Newman, 2001). In Yule-Simon model, the probability of choosing an existing word, which already occurred k times at time t, is $\bar{p}\, k\, \pi(k,t)$, where $\pi(k,t)$ is the fraction of words with frequency k at time t. In order to ascertain whether a PA mechanism might be at work, we construct a running histogram of the frequencies of elements that have been copied, weighting the contribution of each element according to the factor $1/\pi(k,t)$. If this histogram displays a direct proportionality with respect to the frequency k, then one might be observing a PA-driven growth. As an example, we show in Fig. 2.5 the preferential attachment analysis performed on a stream of tags extracted

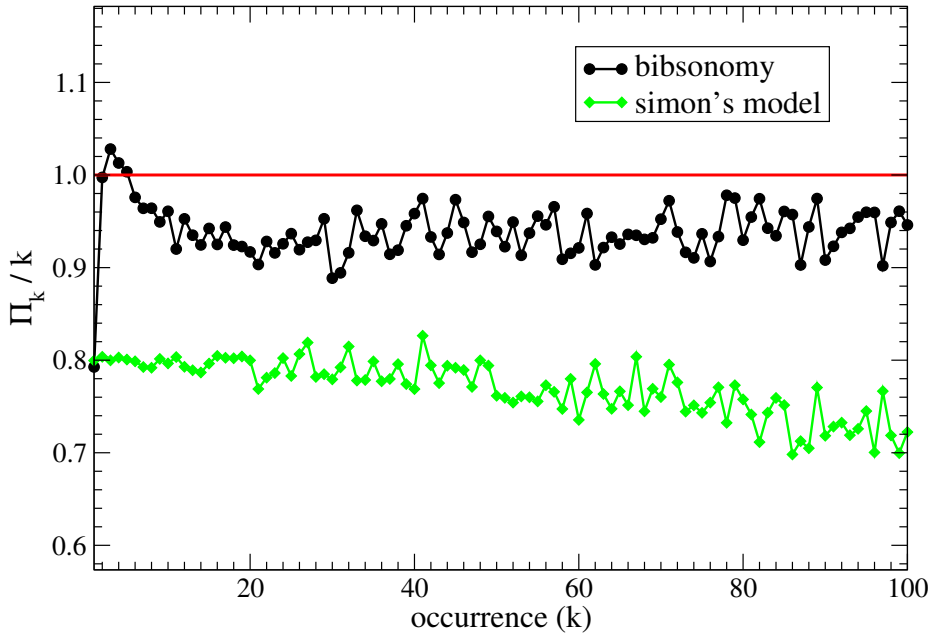

Figure 2.5: Deviations from the preferential attachment rule of a `Bibsonomy` extracted tag stream (circles) as compared with the plain Yule-Simon model (diamonds). The `Bibsonomy` stream contained ca. 1,100,000 tags, while the simulated Yule-Simon model had 1,000,000 tokens and a probability $p = 0.2$ to invent a new token. A straight horizontal line is expected in the case of a preferential attachment mechanism at work. The red line corresponds to $\Pi_k = k$ and is drawn as a guide for the eye. Finite size effects are responsible for the drop at higher frequencies, as extensively discussed in Ref. (Newman, 2001).

from an early stage of `Bibsonomy`, containing ca. 1,100,000 tags in total. The vertical axis shows the above discussed running histogram Π_k, divided the tag occurrence k. In case of preferential attachment, a straight horizontal line is expected.

2.4 Models with memory

In the original Yule-Simon process, the metaphor of text construction is somehow misleading because in that process there is no notion of temporal ordering. All existing words are equivalent and in many respects everything goes as in a Polya urn model (Johnson and Kotz, 1977). However, the notion of temporal ordering may play an important role in determining the dynamics of many real systems. In this perspective it is interesting to investigate models where temporal ordering is explicitly taken into account. A first attempt in this direction has been provided by Dorogovtsev and Mendes (DM) (Dorogovtsev and Mendes, 2000), who studied a generalization of the Barabási-Albert model by introducing a notion of aging for nodes. Each node carries a temporal marker recording its time of arrival into the network, and its probability to be linked to newly added nodes is proportional to its current connectivity weighted by a power-law of its age.

In order to model the observed frequency-rank behavior for the full range of ranking values, we introduce a new version of the "rich-get-richer" Yule-Simon stochastic model (Simon, 1955; Yule, 1925) by enhancing it with a fat-tailed memory kernel. The original model can be described as the construction of a text from scratch. At each discrete time step one word is appended to the text: with probability p the appended word is a new word, never occurred before, while with probability $1-p$ one word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process yields frequency-rank distributions that display

a power-law tail with exponent $\alpha = 1 - p$, lower than the exponents we observe in actual data. This happens because the Yule-Simon process has no notion of "aging", i.e. all positions within the text are regarded as identical.

In our construction we moved from the observation that actual users are exposed in principle to all the tags stored in the system (like in the original Yule-Simon model) but the way in which they choose among them, when tagging a new resource, is far from being uniform in time (see also (Dorogovtsev and Mendes, 2000; Zanette and Montemurro, 2005)). It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones, according to a memory kernel which might be highly skewed. Indeed, recent findings about human activities (Barabasi, 2005a) support the idea that the access pattern to the past of the system should be fat-tailed, suggesting a power-law memory kernel.

We tested this hypothesis with real data extracted from *del.icio.us*: Fig. 1.12 shows the temporal auto-correlation function for the sequence of tags co-occurring with *blog*. Such a sequence is constructed by consecutively appending the tags associated with each post, respecting the temporal order of posts. As explained in the previous Section 1.5, correlations are computed inside three consecutive windows of length T, starting at different times t_w,

$$C(\Delta t, t_w) = \frac{1}{T - \Delta t} \sum_{t=t_w+1}^{t=t_w+T-\Delta t} \delta(\mathsf{tag}(t + \Delta t), \mathsf{tag}(t)) \,,$$

where $\delta(\mathsf{tag}(t + \Delta t), \mathsf{tag}(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t + \Delta t$. From Fig. 1.12 it is apparent that the correlation function is non-stationary over time. Moreover, for each value of the initial time t_w a power-law behavior is observed: $C(\Delta t, t_w) = a(t_w)/(\Delta t + \gamma(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor and $\gamma(t_w)$ is a phenomenological time scale, slowly increasing with the"age" t_w of the system. $c(t_w)$ is the correlation that one would expect in a random sequence of tags distributed according to the frequency-rank distribution $P_{T,t_w}(R)$ pertaining to the relevant data window. Denoting by $R = R\mathsf{max}(T, t_w)$ the number of distinct tags occurring in the window $[t_w, t_w + T]$, we have $c(t_w) = \sum_{R=1}^{R=R\mathsf{max}(T,t_w)} P^2_{T,t_w}(R)$.

Our modification of the Yule-Simon model thus consists in weighting the probability of choosing an existing word (tag) according to a power-law kernel. This hypothesis about the functional form of the memory kernel is also supported by findings in Cognitive Psychology (Anderson, 2000), where power laws of latency and frequency have been shown to model human memory.

More precisely, our model of users' behavior can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of n_0 words. This process is meant to model the behavior of an effective average user in the context identified by a specific tag. At a generic (discrete) time step t, a brand new word may be invented with probability p and appended to the text, while with probability $1 - p$ one word is copied from the existing text, going back in time by x steps with a probability $Q_t(x)$ that decays as a power law, $Q_t(x) = a(t)/(x + \tau)$ (see Fig. 2.6). $a(t)$ is a normalization factor and τ is a characteristic time scale over which recently added words have comparable probabilities.

Fig. 1.4 shows the excellent agreement between the experimental data and the numerical predictions of our Yule-Simon model with long-term memory. Our model, unsurprisingly, also reproduces the temporal correlation behavior observed in real data (see (Cattuto, 2006) for a discussion of this point).

The interpretation of τ (similar to that of the γ parameter introduced above for tag-tag correlations) is related to the number of equivalent top-ranked tags perceived by users as semantically independent. In our model, in fact, 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, i.e. the observed low-rank flattening.

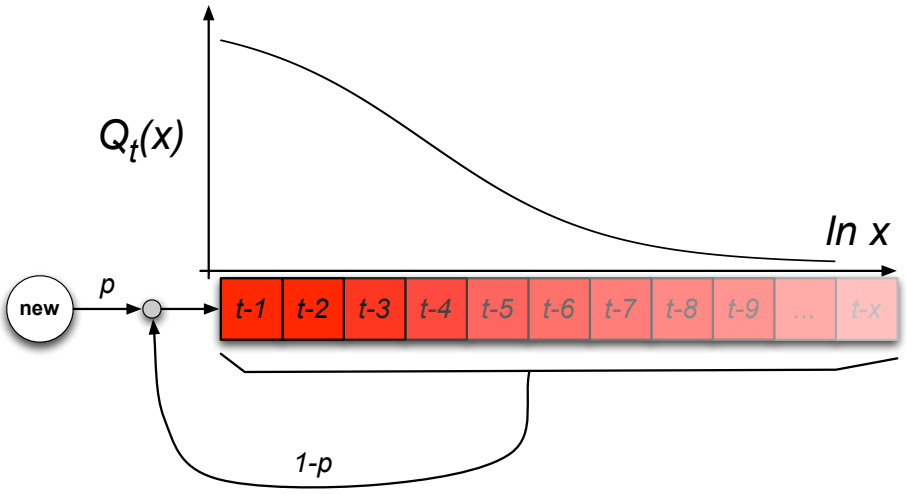

Figure 2.6: Yule-Simon model with fat-tailed memory kernel.

Fitting the parameters of the model, in order to match its predictions (obtained by computer simulation) against the experimental data, we obtain an excellent agreement for all the frequency-rank curves we measured, as shown in Fig. 1.4. This is a clear indication that the tagging behavior embodied in our simple model captures some key features of the tagging activity. The parameter τ controls the number of top-ranked tags which are allowed to co-occur with comparable frequencies, so that it can be interpreted as a measure of the "semantic breadth" of a tag. This picture is consistent with the fact that the fitted value of τ obtained for *blog* (a rather generic tag) is larger than the one needed for *ajax* (a pretty specific one). Additional information on the role of τ as well as that of p in the framework of our model are reported in (Cattuto et al., 2006).

The preferential attachment analysis, presented in the previous Section 2.3.1, reveals how the introduction of a memory kernel changes drastically the properties of the stream. In Figure 2.7 it is shown a comparison with both the Yule-Simon model, which would give a flat curve at 0.6 ($p = 0.4$), and the DM model. Note that our model captures the decay of Π_k/k observable for low k even in the bibsonomy data (see Fig. 2.5).

Our model allows an analytical treatment (Cattuto et al., 2006), at least in the limit case of $\tau = 0$. In this case it is possible to show that the exact analytical expression of the frequency rank distribution $P(R)$ has not a strict power law behavior, even for large R. Instead, it is possible to show that the curve can be approximated with a stretched exponential. The direct comparison between the model simulation and the analytical prediction is shown in Fig. 2.8.

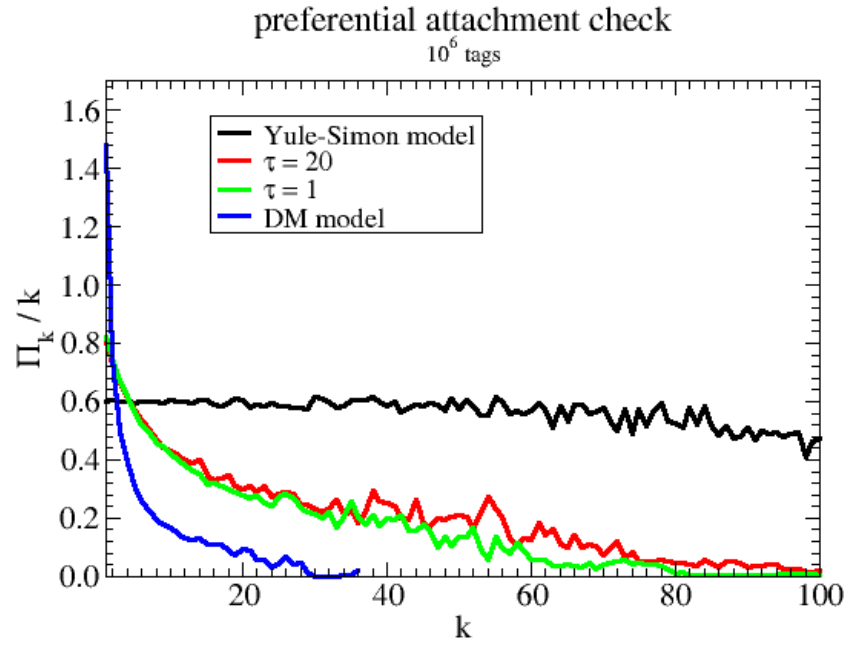

Figure 2.7: Deviations from the preferential attachment rule (Simon's model), in the case of our model and DM model. For all curves, $p = 0.4$ and 10^6 steps were simulated. Finite size effects are responsible for the drop at high frequencies, as extensively discussed in Ref. (Newman, 2001).

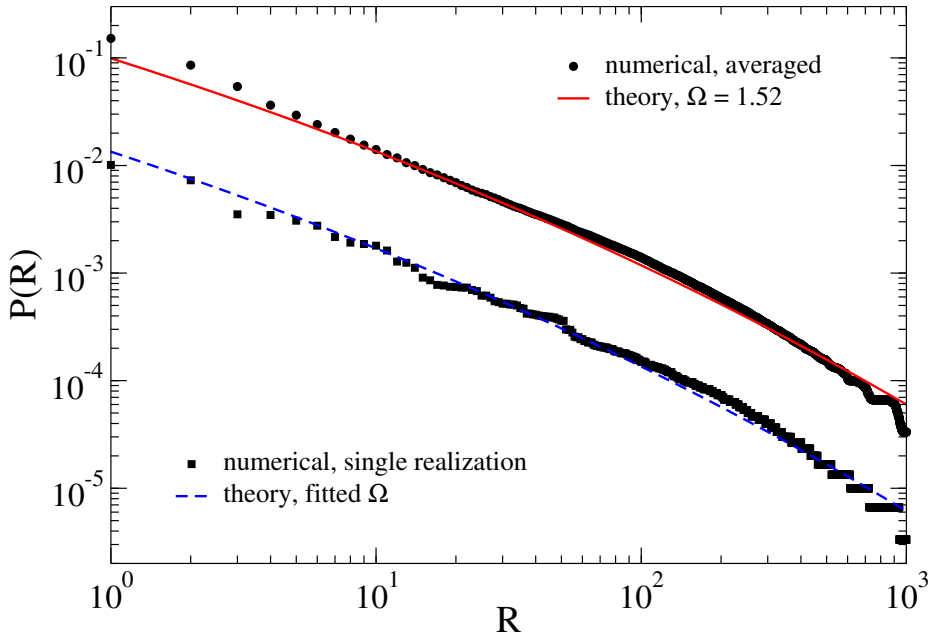

Figure 2.8: Frequency-rank distribution $P(R)$. Numerical data (dots, average over 50 realizations upper curve and a single realization, lower curve) are compared against the analytical prediction (see Eq. 15 in (Cattuto et al., 2006)).

Chapter 3

Experiment I: Semantic analysis on folksonomies

In this chapter we want to list concrete experiments which were performed for analyzing different data sets like Bibsonomy, Delicious and Flickr. For the experiments, several measures performed on streams, defined chapter 1, and on networks, defined in the Deliverable **D3.1**, were used. Thus, this chapter not only offers insights into the data sets collected by the TAGora consortium and general features of tagging systems but it also shows how the previously listed measures interact with each other and how they are used and interpreted.

3.1 Emergence of Patterns in the Tagging Behavior

One of the advantages of tagging systems are that there exist no specific rules how tags have to be used and that there is no predefined set of tags. This helps users to quickly start with tagging and to contribute to tagging systems like Flickr and Del.icio.us. But this unregulated nature of tagging systems can not only be seen as an advantage which lowers the entry barriers for new users but it also partially reduces the potential benefits from such a system. For example, there exist no rules how to handle compound words like “San Francisco” or “garbage can” or that e.g. the singular form of nouns should be preferred over the plural form. Without such rules, an increased amount of work is required during the search for resources in order to get a high recall. For example, should the user search for “sanfrancisco”, “san_francisco” or for “san” and “francisco”?

It can be expected that mechanisms similar to the naming or guessing games (see (TAGora, 2007)) may not only lead to the development of a common vocabulary but also to the emergence of patterns in the user behavior how they deal with the above mentioned deficiencies of tagging systems. The driving force behind the expected development of patterns would be the benefit during searching resources. It would be comparable to the attempt of the agents in the naming game to increase the number of rounds where speaker and hearer use the same name for an object.

Analyzing the emergence of such common tagging behavior would show that there is an influence of the users on each other which not only affects the local behavior while tagging a single resource but that it also leads to a permanent learning effect and change of behavior. It depends on the existence of such permanent learning effects in how far tagging systems may remain unregulated as they currently are or whether mechanisms have to be found which allow to deal with the inconsistencies in the tagging behavior of the users. For example, one may try to identify and merge singular and plural forms of nouns or the different spellings of compound words.

If the emergence of such patterns in the tagging behavior can be shown it would be very likely that they are caused by learning effects inherent to tagging systems and not learning processes or changes taking place in the real world. For example, when the emergence of new tags are

$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$
31.394	1.372.103	18.778.597	82.296.035

Table 3.1: Size of the used Delicious dataset.

observed which take over already established tags for a certain resource (like it is shown in (Steels, 2006)) it is not sure to which extent this observed adaptation of user behavior is caused by changes in the world or by learning effects inherent to the tagging system.

Furthermore, it has to be shown that the learning effect is not only restricted to a single resource (e.g. that "san_francisco" dominates "sanfrancisco") but that there is a global tendency to one of the possible choices. Otherwise, the observed adaption process of the users may not be permanent but correspond to an adaption process of the language system to the "dialogue" with the other users which already tagged the resource. The situational adaptation of the language system is a phenomenon which is also observable in real life where it helps to optimize the communicative success and minimize cognitive effort (cf. (Steels, 2006)).

Two experiments were performed where the emergence of patterns in the tagging behavior of users were analyzed. In the first experiment, it was analyzed whether a specific kind of handling compound words was preferred by the users and in the second experiment whether the singular form of a noun is preferred over the plural form. Both experiments were carried out on the Delicious data set collected by the TAGora consortium. In Tab. 3.1 one can see how many tags, tag assignments, resources and users are contained in the dataset.

3.1.1 Dealing with Compound Words

A typically problem for users of Delicious is that it isn't allowed to use a space character in the tag names. Thus, it isn't possible to write a compound word like "San Francisco" as a single tag if no mechanism is found to circumvent this problem. There exist several ways how this can be done. Six typical variants for dealing with compound words can be observed in the Delicious dataset:

- **Compound** For each of the compounds a separate tag is assigned to the resource (e.g. "san" and "francisco").
- **Hyphen** The compounds are separated by a hyphen (e.g. "san-francisco").
- **Underscore** The compounds are separated by an underscore (e.g. "san_francisco").
- **Plus** The compounds are separated by a plus (e.g. "san+francisco").
- **Dot** The compounds are separated by a dot (e.g. "san.francisco").
- **Concat** The compounds are concatenated to a single word (e.g. "sanfrancisco").

In the following, we will analyze in how far one of this six variants for dealing with compound words is preferred by the users and whether over time one of the variants gains more popularity than the other. For this purpose, we analyzed the cumulated tag occurrence for the six variants for 54.936 different compound nouns which were taken from the list of nouns in Wordnet 3.0[1]. For each of the compound nouns from Wordnet, we created the six variants described above. Additionally, for each compound noun also its plural form was created in the six variants. Subsequently, the Delicious dataset was reduced to the tags which are contained in the list of generated compound variants and their tag assignments. The resulting number of distinct tags and tag assignments are available in Tab. 3.2.

[1] http://wordnet.princeton.edu/

$\|T\|$	$\|Y\|$
45.175	44.351.838

Table 3.2: Size of the Delicious dataset reduced to compound nouns.

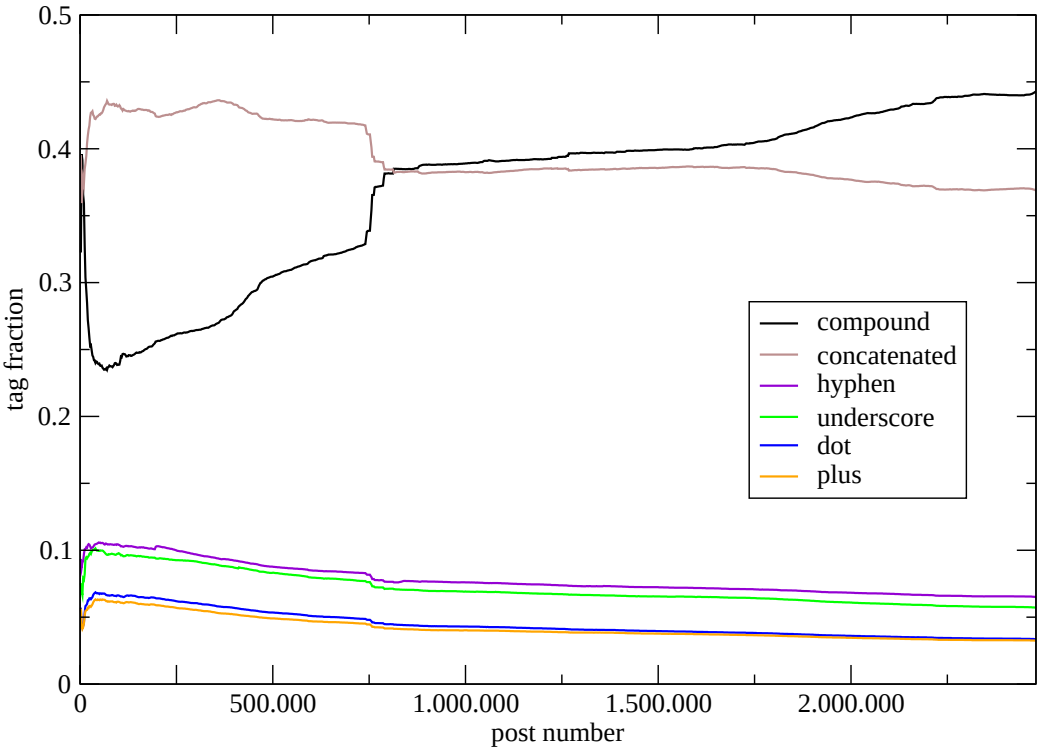

Figure 3.1: Cumulated occurrences for the variants for expressing compound words in Delicious shown as a function of time, measured in number of posts.

For the reduced Delicious dataset we accumulated the number of tag assignments for each of the six variants and plotted it as a function over time. For the variants consisting of a single tag, counting their occurrences is trivial. For the variant where the compound word was split into two distinct tags it was counted as one occurrence of this variant if the first word and the second word of the compound occur in the same posting, e.g. if “san” and “francisco” are assigned in the same posting.

Fig. 3.1 shows the results for the different variants. One can see that the variant where the compound words are simply concatenated (e.g. “sanfrancisco”) and the variant where more than one tag is used for the compound are by far the most popular variants. At the beginning, concatenating a compound word into a single tag was the most popular variant which remains at a steady value around 43% but using more than one tag for a compound became increased its popularity during that time, mainly on the cost of the other four, less popular variants.

Around post number 750.000 one can see a sudden peak in the popularity of using separate tags where the fraction of this variant is increased from approximately 33% to 38%. Such a sudden change in the fractions points to external influence factors which lead to a preference of the variant with separate tags. Possible explanations might be changes to the tagging interface (e.g. the space became the separator for tags and thus the users automatically split compound words into several tags) or that using separate tags was propagated in blogs etc. as the preferred way of handling compound words.

After this sudden peak one can observe that the fractions of the two most popular variants remain quite stable around the same value. The variant of using separate tags gains slightly more pop-

$\|T\|$	$\|Y\|$
55.423	54.171.476

Table 3.3: Size of the Delicious dataset reduced to tags which are singular or plural forms of nouns.

ularity on the cost of the four less popular variants. Around post number 1.750.000 a new trend can be observed where an increase of the popularity of the separate tags variant also results in a decrease of the concatenated word variant. This trend seems to slow down at the end of the period covered by our Delicious dataset but it didn't come to a stop.

This clearly shows that the users in a tagging system influence each other enough so that a permanent change of their behavior can be reached which is not restricted to the tagging of a single resource. Furthermore, by analyzing such a tagging system specific behavior, we can exclude an influence from changes to the real world. Nevertheless, the fast gain of popularity for the two tag variant in a very short period shows that there is still an important influence of other factors than that of processes similar to naming games.

For the future, we plan to analyze in how far similar patterns in the user's tagging behavior can be observed in other tagging systems. Of special interest is whether also a sudden peak of popularity can be observed in other systems around the same time. This would mean that it isn't caused by Delicious specific influence factors (e.g. its tagging interface) but by external factors like propagating a certain variant in blogs etc.

3.1.2 Preference of Certain Flexion Forms

A further experiment was performed where it was analyzed whether the users in a tagging system prefer either the singular or plural form of a noun or whether no preference can be observed. Furthermore, also the development of the distribution between singular and plural forms was of interest. Again, the above described Delicious dataset was analyzed. The setup of the experiment was very similar to that described in section 3.1.1: We took from Wordnet 3.0 the complete list of nouns and created for each of the nouns its plural form. The result was a list of 132.545 nouns in their singular and plural form. Subsequently, we reduced the set of tags to those contained in the noun list and only retained the corresponding tag assignments. The resulting number of distinct tags and tag assignments are available in Tab. 3.3.

In Fig. 3.2 one can the distribution between singular and plural forms of nouns. In opposite to the results for the usage of different compound variants in section 3.1.1, one can not see any change in the distribution between singular and plural forms of nouns. The fraction of singular forms of nouns constantly lies around 83%, i.e. no permanent change in the users' global tagging behavior can be observed. This may be because the problem of using different flexion forms is not as obvious as the handling of compound words. In the case of the compound words, the users are forced by the tagging system to find a way how to deal with them while the usage of different flexion forms may be unnoticed by the users. Thus, they are not searching for a solution and subsequently do not look how other users deal with the problem.

3.2 Comparing Vocabulary Usage and Richness

In this section, we want to analyze in how far one can observe different vocabularies for different tagging systems, i.e. whether they are focused on different topics. For example, by looking at Flickr and Delicious one can see that in Flickr many photos are from vacations while in Delicious many bookmarks are related to technology.

For the experiment, we used the noun categories provided by Wordnet 3.0. Each noun has as-

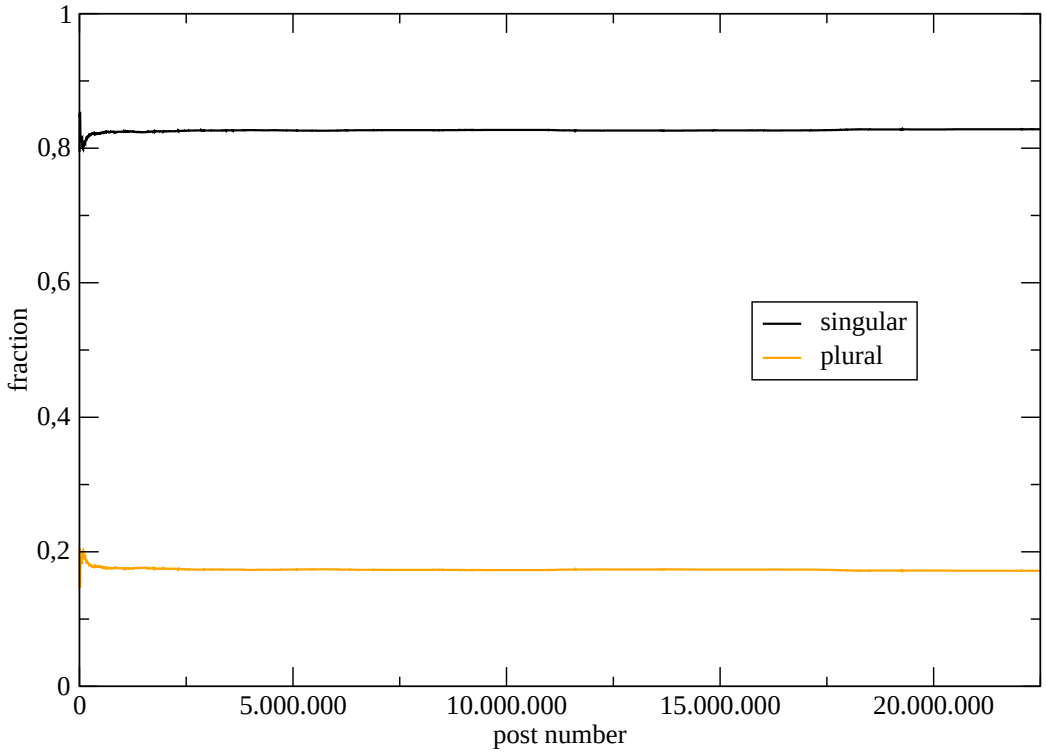

Figure 3.2: Cumulated occurrences for the singular and plural forms of tags in Delicious shown as a function of time, measured in number of posts.

signed at least one of the categories, which can roughly be mapped to the above mentioned topics. For example, pictures from the vacation or last night's party will have assigned many tags from the location or the person category. The following list contains the most important categories for the following analysis[2]:

- **Communication**: Nouns denoting communicative processes and contents (e.g. *hypothesis, lemma, mass medium, comic strip, life*).
- **Person**: Nouns denoting people (e.g. *alumnus, brother, djinny, abraham lincoln, alan turing*).
- **Location**: Nouns denoting spatial positions (e.g. *african nation, abu dhabi, acre*).
- **Plant**: Nouns denoting plants (e.g. *ginko, fungus, leaf, philodendron*)
- **Act**: Nouns denoting acts or actions (e.g. *achievement, analysis, intermezzo, dialysis*).
- **Object**: Nouns denoting natural objects (not man-made) (e.g. *africa, nebula, volcano*).
- **Cognition**: Nouns denoting cognitive processes and contents (e.g. *aim, formula, topos*).
- **Artifact**: Nouns denoting man-made objects (e.g. *aquarium, bookshelf, fresco, knife*).

During our experiments we analyzed the vocabulary richness, i.e. the number of distinct tags belonging to one of the Wordnet categories, and the vocabulary usage, i.e. how often was a tag from one of the categories assigned to a resource. We did this for a Flickr and a Delicious dataset. The Delicious dataset is the same as in section 3.1. For the analysis of Flickr we used a subset

[2]A complete list of the noun categories in WordNet is available at `http://wordnet.princeton.edu/man/lexnames.5WN.html`

	$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$
Flickr	142.939	665.558	7.295.296	34.228.897
Delicious	31.394	1.372.103	18.778.597	82.296.035

Table 3.4: Size of the unfiltered Flickr and Delicious datasets.

	$\|T\|$	$\|Y\|$
Flickr	50.200	20.535.583
Delicious	55.423	54.171.476

Table 3.5: Size of the Flickr and Delicious dataset reduced to nouns.

of the Flickr dataset collected by the TAGora consortium. It covers resources uploaded between January 2004 and September 2005. The statistics of the unfiltered Flickr and Delicious dataset are available in Tab. 3.4.

As it was said, the analysis was based on the list of nouns available in Wordnet, i.e. 117.798 distinct nouns. For each of the nouns we also created the plural form. Compound words were concatenated to a single string. This resulted in 232.498 distinct strings against which the tags from the Flickr and Delicious dataset were matched and subsequently assigned to the noun categories. In Tab. 3.5 one can see the size of the two datasets reduced to the tags contained in the list of nouns.

In Fig. 3.3 and 3.4 one can see the results for the Flickr and Delicious dataset. By looking at the number of tag assignments for each of the noun categories (Fig. 3.3), one can see that in Flickr the tag assignments belonging to the person and location category have a higher importance than in Delicious. In the Delicious dataset the tag assignments belonging to the communication and cognition category are more important than in Flickr. This is not so surprising as it reflects the above mentioned focus of e.g. Flickr on photos from the last vacation or party.

It is more interesting that the differing importance of the categories on the level of tag assignments is not reflected on the level of distinct tags. For example, in Fig. 3.4 one can see that there are only minor differences in the relative numbers of distinct tags for the communication, person, location and cognition category. Also for the other noun categories only minor differences in the size of the vocabulary can be observed.

The only exceptions are the plant and the animal category. For both categories, the Flickr vocabulary contains (relative to all distinct tags) twice as much tags as the corresponding vocabulary in Delicious (see Fig. 3.4). But concluding from the findings e.g. for the person or location category, this may only be partially caused by the higher usage of tags from the plant category in Flickr. For the latter two categories, a similar difference in the vocabulary usage only led to minor differences in the size of the vocabulary.

The special role of the plant and the animal category in the Flickr dataset also becomes obvious by directly comparing the relative size of the vocabulary with its relative usage, i.e. the number of tag assignments (see Fig. 3.5). One can see that the size of the plant and animal vocabulary in Flickr is significantly larger than one may conclude from its number of tag assignments.

In the previous description of the experiment we identified two different patterns in the relation between the size of a vocabulary and the number of corresponding tag assignments which can be used as an indicator for an increased importance of a topic or rather the existence of a community of users. On the one hand we had the pattern where the size of the vocabulary remains quite stable and the increased importance is only reflected on the level of tag assignments. This pattern is characteristic for a very large community (e.g. people sharing their photos from their last vacation). On the other hand we had the pattern where the size of the vocabulary is very large compared to the importance based on the number of tag assignments. This pattern is characteristic for a small but specialized community (e.g. people sharing plant photos). In the future, we plan to explore in

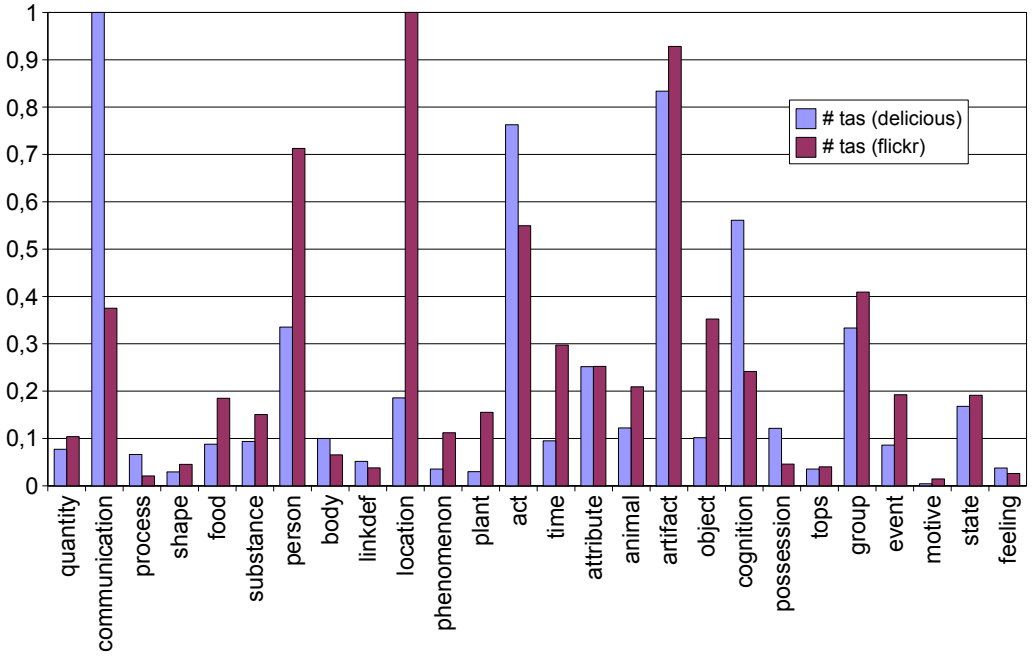

Figure 3.3: Number of tag assignments for the different noun categories in Flickr and Delicious. The values are normalized, i.e. the value is relative to the number of tag assignments in the most often used category.

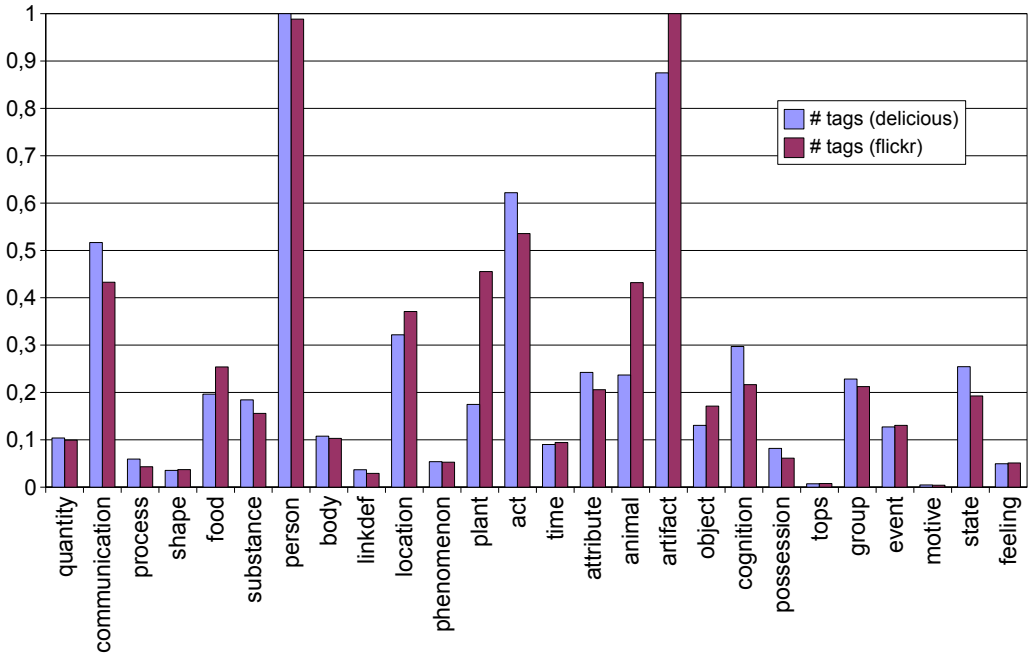

Figure 3.4: Number of distinct tags for the different noun categories in Flickr and Delicious. The values are normalized, i.e. the value is relative to the number of distinct tags in the most often used category.

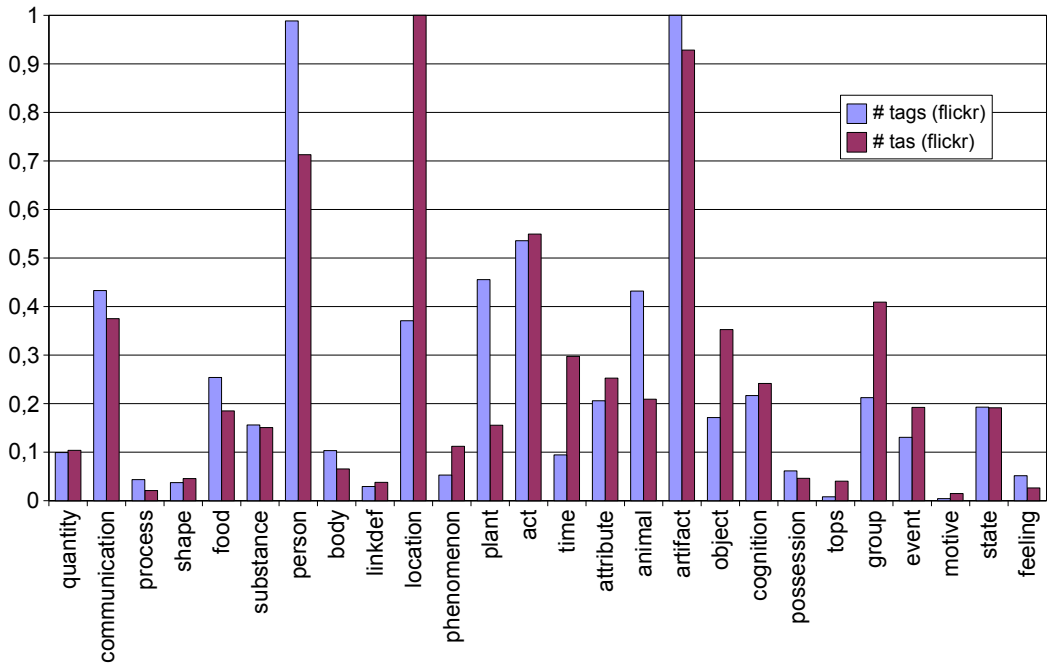

Figure 3.5: Comparison between the relative size of the vocabulary and the relative number of tag assignments in Flickr.

how far the analysis of vocabulary size and usage are suitable for detecting user communities in tagging systems.

Chapter 4

Experiment II: Folksonomy aided recommender systems

While the Semantic Web has evolved to support the meaningful exchange of heterogeneous data through shared and controlled conceptualisations, Web 2.0 has demonstrated that large-scale community tagging sites can enrich the semantic web with readily accessible and valuable knowledge. In this section, we investigate the integration of a movies folksonomy with a semantic knowledge base about user-movie rentals. The folksonomy is used to enrich the knowledge base with descriptions and categorisations of movie titles, and user interests and opinions. Using tags harvested from the Internet Movie Database, and movie rating data gathered by Netflix, we perform experiments to investigate the question that folksonomy-generated movie tag-clouds can be used to construct better user profiles that reflect a user's level of interest in different kinds of movies, and therefore, provide a basis for prediction of their rating for a previously unseen movie and improve recommender systems.

4.1 Recommender Systems

Recommender systems are usually used in one of two contexts: (1) to help users locate items of interest they have not previously encountered, (2) to judge the degree of interest a user will have in item they have not rated. With the growing popularity of on-line shopping, E-commerce recommender systems (Schafer et al., 1999) have matured into a fundamental technology to support the dissemination of goods and services. Much research has been undertaken to classify different recommendation strategies (Burke, 2002; Herlocker et al., 2004), but here we divide them broadly into two categories.

Collaborative recommendation is probably the most widely used and extensively studied technique that is founded on one simple premise: if user A is interested in items w, x, and y, and user B is interested in items w, x, y, and z, then it is likely that user A will also be interested in item z. In a collaborative recommender system, the ratings a user assigns to items is used to measure their commonality with other users who have also rated the same items The degree of interest for an unseen item can be deduced for a particular user by examining the ratings of their neighbours. It has been recognised that users interest may change over time, so time-based discounting methods have been developed (Billsus and Pazzani, 2000; Schwab et al., 2001) to reflect changing interests.

Content-based recommendation represents the culmination of efforts by the information retrieval and knowledge representation communities. A set of attributes for the items in the system is conceived, such as the keywords and term frequencies for documents in a repository, so the system can build a profile for each user based on the attributes present in the items that user has rated highly. The interest a user will have in an unrated item can then be deduced by calculating its similarity to their profile based on the attributes assigned to the item.

Such systems are not without their deficiencies, the most prominent of which arise when new items and new users are added to the system - commonly referred to as the *ramp-up* problem (Konstan et al., 1998). Since both content-based and collaborative recommender systems rely on ratings to build a user's profile of interest, new users with no ratings have neutral profiles. When new items are added to a collaborative recommender system, they will not be recommended until some users have rated them. Collaborative systems also depend on the overlap in ratings across users and perform badly when ratings are sparse (i.e. few users have rated the same items) because it is hard to find similar neighbours.

Hybrid recommender systems, i.e. those which make use of collaborative and content based approaches, have been developed to overcome some of these problems. For example, collaborative recommender systems do not perform well with respect to items that have not been rated, but content-based methods can be used to understand their relationship to other items. Hence, a mixture of the two approaches can be used to provide more robust systems. More recent recommender systems have also investigated the use of ontologies to represent user profiles (Middleton et al., 2004). Benefits of this approach are more intuitive profile visualisation and the discovery of interests through inferencing mechanisms.

4.2 Recommender Architecture

To gather the information necessary to construct profiles that describe the kinds of movies a user is interested in, we combine data harvested from two sources: a collaborative a folksonomy, i.e. a popular movie collaborative tagging systems, and a data base of user-movie rentals.

4.2.1 Data Sources

For movie tagging data, we make use of the Internet Movie Database (IMDB) (The Internet Movie Database , IMDB); an online database containing extensive information on movies, actors, television shows, and production personnel. IMDBholds information on approximately 960,000 titles and 2,300,000 people, and is the largest known accumulation of data about films (The Internet Movie Database - Wikipedia Entry). In terms of tagging, IMDBallows users to add *keywords* to titles to describe arbitrary features of the movie. Typically, these are used to denote important scenes in the film (e.g. `sword-fight`, `kidnapping`, `car-chase`), plot themes (e.g. `love`, `revenge`, `time-travel`), locations (e.g. `space`, `california`), film genres (e.g. `independent-film`, `non-fiction`, `cult-favorite`), and background data (e.g. `based-on-novel`, `based-on-true-story`). On average, a popular movie has between 50 and 150 keywords attached to it.

Currently, IMDBuses this tagging data to create a movie search tool that helps users to find popular movies based on their keywords. A screen shot of this interface in shown in Figure 4.1 and contains two panels: on the left, a tag cloud is used to display keywords; and on the right, a list of the top movies that contain the currently viewed keywords. In this particular example, the keywords `space` and `android` are used as the search terms.

Note that IMDBis a tagging *free-for-all*: users may tag any resources. However the addition of keywords to a movie is moderated, but this is used mainly to prevent spam attacks and not to manage the keywords used. When adding keywords to a movie, users can see the keywords that have already been added, but the individual keyword assignments by each user cannot be seen. Instead, a simple list of keywords is maintained for each movie and duplicates are not allowed.

The other source of information, the user rentals data, is provided by Netflix (Netflix Homepage) as part of the Netflix Prize (Netflix Prize Homepage). Netflix is an online DVD rental service, established in 1998, the provides a flat rate, mail-based, rental service to customers in the United States. Their current DVD collection contains around 75,000 titles, offered to a customer base of

Figure 4.1: A screen shot of the IMDB keyword search interface.

over 6 million individuals. After renting a movie, customers may enter their rating of the movie into the Netflix database via the website, using a discrete score from 1 to 5.

In October 2006, Netflix began a competition to find better recommendation systems, offering a grand prize of $1 million to anyone managing to improve on their own algorithm by 10%. To drive this competition, Netflix published a large set of movie rating data from their database featuring 480,189 customers and 100,480,507 ratings across 17,770 movie titles.

4.3 Recommendation Method

To explore the relationship between the way a user rates movies and the keywords that are assigned to movies, we have devised two prediction algorithms that guess the rating a user would give to a previously unrated movie based on tag-clouds that depict their interests. For comparison, we also specify a naive average-rating algorithm were the average rating for a movie across all users is used as the predicted rating.

4.3.1 Notation

Let us denote a given user by $u \in U$, where U is the set of all users, a movie by $m \in M$, where M is the set of all available movies, and a rating value by the integer $r \in \{1, 2, 3, 4, 5\} \equiv R$. We indicate the set of movies rated by user u as M_u. On this set we define the rating function for user u as $f_u : m \in M_u \mapsto f_u(m) \in R$.

When keywords or tags are available for a movie m, we denote by K the global set of keywords, by K_m the set of keywords (or tags) associated with movie m, and by N_k the global frequency of occurrence of keyword k for all movies. We can then introduce a notion of *rating tag-cloud* $T_{u,r}$ for a given user u and rating r as the set of couples (k, n_k), where $k \in K$ indicates a keyword (or tag) and $n_k = n_k(u, r)$ is its frequency of occurrence for all movies that user u has associated with rating r. That is,

$$n_k(u, r) = |\{m \in M_u \,|\, k \in K_m \ \wedge \ f_u(m) = r\}| \ . \tag{4.1}$$

Two sample rating tag-clouds are shown in Figure 4.2; the left one is a rating 1 tag-cloud, and the right one is a rating 5 tag-cloud. The size of keywords is proportional to the logarithm of their frequency of occurrence in the tag-cloud they belong to.

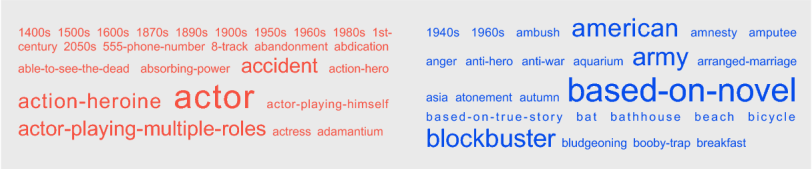

Figure 4.2: Sample rating tag-clouds (left: rating 1, right: rating 5).

4.3.2 Average-based Rating

A very simple rating prediction strategy can be implemented by assuming that a given user u^* will rate a new movie m^* ($m^* \notin M_{u^*}$) according to the average rating that the movie received by all other users. We compute the average rating of movie m as

$$\bar{r}_m = \frac{1}{|U_m|} \sum_{u \in U_m} f_u(m)\,, \tag{4.2}$$

where $U_m = u \in U | m \in M_u$ is the set of users that have rated movie m, and $|U_m|$ is its cardinality. In this scheme, the predicted rating for movie m^* is the integer $r^* \in R$ that is nearest to $\bar{r}_{m^*}$.

4.3.3 Simple Tag-Cloud Comparison

In this scheme we guess the rating that user u^* would give to movie m^* by comparing the set of keywords K_{m^*} associated with the movie against the rating tag-clouds $T_{u^*,r}$ of user u^* for different ratings. We guess the rating r^* as the one corresponding to the tag-cloud (of user u^*) that most closely resembles the set of keywords K_{m^*}, as measured by the number of keywords that K_{m^*} shares with the tag-clouds of user u^* for different ratings:

$$\sigma(u^*, m^*, r) = |\{(k, n_k) \in T_{u^*,r} \,|\, k \in K_{m^*}\}|\,. \tag{4.3}$$

4.3.4 Weighted Tag-Cloud Comparison

In this hybrid scheme we try to take into account weights both at the keyword level (through their frequencies n_k) and at the tag-cloud level, though a measure of tag-cloud similarity. Given a new (in the sense of unrated) movie m^*, we consider the set of keywords K_{m^*} and introduce a notion of "similarity" between K_{m^*} and a given tag-cloud $T_{u,r}$. We define such a measure of similarity as:

$$\sigma(u, m, r) = \sum_{\{(k,n_k) \in T_{u,r} \,|\, k \in K_m\}} \frac{n_k}{\log(N_k)}\,, \tag{4.4}$$

that is we sum over all keywords which K_{m^*} and the tag-cloud $T_{u,r}$ have in common, and we weight each keyword k proportionally to its frequency n_k in the tag-cloud, and inversely proportional to the logarithm of its *global* frequency N_k, as commonly done in TFIDF term-weighting schemes.
We subsequently define the weighted average rating as

$$\bar{\sigma}(u, m) = \frac{1}{S(u, m)} \sum_{r \in R} r\, \sigma(u, m, r)\,, \tag{4.5}$$

where $S(u, m) = \sum_{r \in R} \sigma(u, m, r)$ is a normalization factor. Thus, $\bar{\sigma}(u, m)$ is an estimate of a user rating based on the weighted similarity between the set of movie keywords and the user's rating tagclouds (themselves weighted). This information can be used by itself, to guess a user rating, or it can be used to improve a prediction based on other techniques.

In our experiment we decided to use the rating $\bar{\sigma}(u, m)$, estimated from the tag-cloud similarity, to improve the simple rating estimate based on the per-movie average rating (see section 4.3.2). We combine the two estimates by computing their weighted average. That is, given a user u^* and a movie m^*, our estimate for the rating is

$$\sigma^*(u^*, m^*) = (1 - \gamma)\, \bar{r}_{m^*} + \gamma\, \bar{\sigma}(u^*, m^*)\,, \tag{4.6}$$

where $0 < \gamma < 1$ is a factor weighting the contribution of the two estimates. In our experiment we set $\gamma = 1/2$. We guess the rating r^* as the integer in R that lies closest to the weighted average $\sigma^*(u^*, m^*)$.

Of course, the above strategy can only be used when the set of keywords K_{m^*} associated with movie m^* is non-empty. If K_{m^*} is empty our implementation resorts to using the simple strategy of section 4.3.2 (equivalent to setting $\gamma = 0$ in Eq. 4.6).

4.4 Experiment and Results

To test the algorithms presented, we extract a training set from the full Netflix data dump containing the ratings of 500 randomly chosen users. For each user, a test set made up from their last 100 ratings is removed from the training set so the accuracy of our algorithms can be tested. For each user, the root mean squared error (RMSE) is recorded, along with the percentage of exactly matched ratings. Given a set of predicted ratings $\{r_i\}$ and the corresponding set of actual ratings $\{r_i^*\}$, the RMSEis defined as:

$$\text{RMSE}(\{r_i\}, \{r_i^*\}) = \sqrt{\frac{1}{N} \sum_i (r_i - r_i^*)^2}\,. \tag{4.7}$$

A summary of the results follows:

	Average Rating	Unweighted	Weighted
Correct	36.12%	44.15%	42.47%
Incorrect	63.99%	55.85%	57.53%
RMSE	1,131	1.074	0.961

The unweighted tag-cloud comparison does perform better than the naive average rating, with a moderate increase in the percentage of correctly rated movies. Using the weighted tag cloud comparison improves the RMSE, but with a slight drop in the fraction of exactly matched ratings. Figure 4.3 contains two scatter plots (unweighted and weighted tag-cloud comparison techniques) showing the RMSEfor each user against the number of movies in their training set. These plots show two interesting features: (i) the weighted comparison technique has a smaller error range than the unweighted comparison (ii) the error rate seems to be independent of the number of movies rated. To visualise the distribution of predicted ratings for each of the algorithms, we present two histograms in Figure 4.4: one showing the distributions of the predicted ratings, and one showing the global distribution of actual ratings. From these charts, it is clear that the rating categories 1 and 2 are being neglected.

In order to gain more insight into the behavior of our prediction schemes, we study the distribution of predicted ratings as a function of the actual rating. Fig. 4.5 shows the (color-coded) probability distribution of predicted ratings as a function of the actual movie rating, for the simple average-based scheme (left figure) and the weighted tag-cloud comparison scheme (right figure).

A perfect prediction scheme would appear as a unity matrix, with ones along the main diagonal and zeros elsewhere. Fig. 4.5 shows that both prediction schemes behave poorly for low (1 and

2) and high (5) values of the actual rating, as both schemes predict intermediate ratings (3 and 4) with high probability, independent of the actual rating (bright rows in the plots).

We observe that the weighted tag-cloud scheme provides enhanced contrast throughout the rating range. For intermediate values of the actual rating (3 and 4) it improves significantly over the average-based scheme, with a better separation of the diagonal elements (3-3 and 4-4, correct predictions) over the off-diagonal ones, in particular over the elements corresponding to the incorrect predictions 3-4 and 4-3. For the highest actual rating (5) the weighted tag-cloud scheme features a distribution of predicted values which is more skewed towards high ratings, but on average it still fails to predict the correct rating. The same happens for low actual ratings (1 and 2), where the weighted tag-cloud scheme displays a distribution of predicted values which is more skewed towards low-values, but still fails to predict 1s and 2s with a significant probability.

In terms of future work, this evaluation shows that intermediate ratings are predicted rather well, and additional work is needed to make better prediction of extreme rating values, both high and low.

4.5 Conclusions and Future Work

We have demonstrated that a movie recommendation system can be built purely on the keywords assigned to movie titles via collaborative tagging. By building different tag-clouds that express a user's degree of interest, a prediction for a previously unrated movie can be made based on the similarity of its keywords to those of the user's rating tag-clouds. With further work, we believe our recommendation algorithms can be improved by combining them with more traditional content-based recommender strategies. Since IMDBprovides extensive information on the actors, directors, and writers of movies, as well as demographic breakdowns of the ratings, a more detailed profile can be constructed for each user. Also, our recommendation algorithms have not exploited any collaborative recommender techniques. Further research may show that rating tag-clouds are a useful and more efficient way to find neighbours with similar tastes.

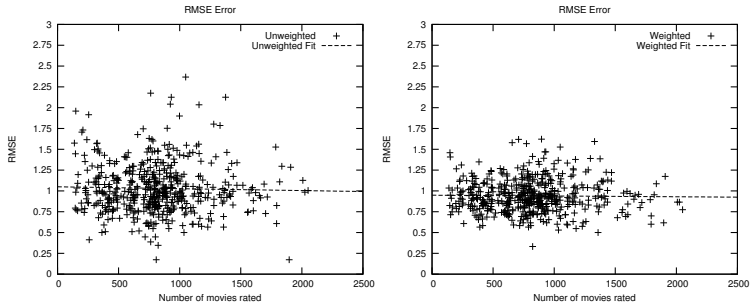

Figure 4.3: Scatter plots to show the level of accuracy for each rating technique in terms of the number of movies rated by the user.

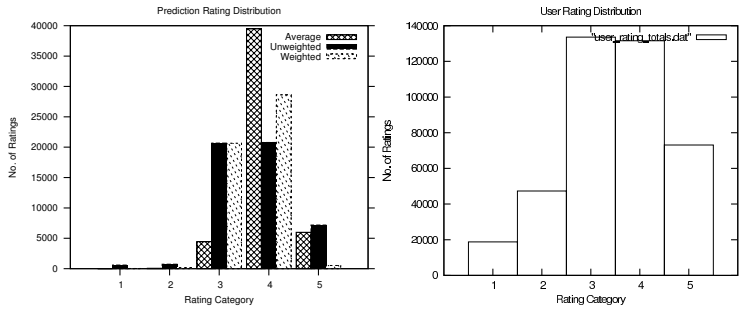

Figure 4.4: Histograms showing the number of predictions made in each rating category, and the overall rating distribution

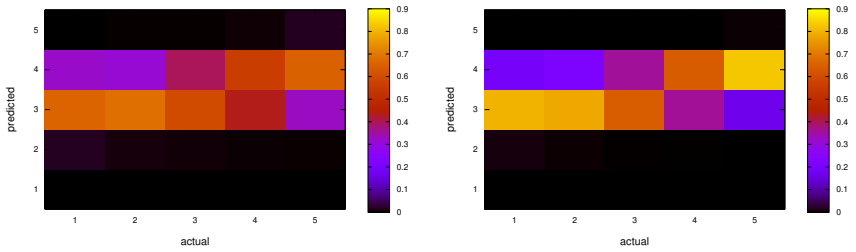

Figure 4.5: Distribution of predicted ratings as a function of actual movie rating, for the simple average-based scheme (left plot) and the weighted tag-cloud comparison scheme (right plot). For each value of the actual rating (horizontal axis), a normalized histogram of the predicted ratings (vertical axis) was built, displaying how predicted values are distributed. Because of normalization, the sum of values along all columns is 1.

Chapter 5

Conclusions and Perspectives

5.1 Conclusions

In the present report, we introduce the study of folksonomies in their stream view. In other words, we focus on the continuous flow of metadata entered into the system by users. This global stream of tag assignments reveals some striking statistical regularities and features.

We describe relevant statistical measures and review known results about text streams, notably a subject of computational linguistics. Particular attention has been devoted in recent years to the frequency distribution of words and to the vocabulary size of texts and corpora.

In the case of a folksonomy we consider these and other quantities, such as stream correlations or measures of “preferential attachment” borrowed by complex networks theory. In particular we report a systematic investigation on vocabulary growth in folksonomies, i.e. how the number of different tags present in the system evolves with time, both globally and restricted to a variety of contexts. Beyond the pure theoretical interest, understanding vocabulary growth is crucial in order to control the scalability and the effectiveness of folksonomies.

In order to explain the statistical observations, a number of theoretical problems arise. To this aim, we address the possibility to introduce simple stochastic models capturing the microscopic mechanism at the base of user tagging activity. After a brief selected review of stochastic models introduced for explaining the word frequency distribution in texts, we present an original stochastic model which recover many observed specific features of folksonomies. However, more ingredients are needed in order to improve the models: both the introduction of more cognitive based user behavior, as also the adoption of a multi-agent paradigm will surely be addressed in the next years.

We accompany the first attempt at statistical and theoretical approach with two experiments aimed at bridging the gap between collaborative tagging systems and the semantic web approach. In the first experiment, we perform an analysis of the commonly used lexical forms of tags, with the goal of exploring how much such vocabulary analysis is suitable for detecting user communities in tagging systems.

In a second, more applied experiment we try to exploit data taken from a collaborative tagging systems to improve the automatic recommendation strategy in a commercial context, the Netflix challenge. This study represent a first step in the contribution of the TAGora project to improving navigability and control strategies of collaborative tagging systems.

Bibliography

Project gutenberg. URL http://promo.net/pg.

John R. Anderson. *Cognitive Psychology and Its Implications*. Worth Publishers, New York, 5th edition edition, 2000.

S. Naranan V. K. Balasubrahmanyan. Quantitative linguistics and complex system studies. *Jorunal of Quantitative Linguistics*, 3:177–228, 1996.

V. K. Balasubrahmanyan and S. Naranan. Algorithmic information, complexity and zipf's law. *Glottometrics*, 4:1–26, 2002.

A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005a.

Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005b. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0505371.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/9910332.

D. Billsus and M. Pazzani. User modeling for adaptive news access, 2000. URL http://citeseer.ist.psu.edu/billsus00user.html.

Stefan Bornholdt and Holger Ebel. World wide web scaling exponent from simon's 1955 model. *Phys. Rev. E*, 64(3):035104, Aug 2001. doi: 10.1103/PhysRevE.64.035104. e-print cond-mat/0008465.

Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002. ISSN 0924-1868.

C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. 2007a. URL http://arxiv.org/abs/0704.3316.

Ciro Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46(0):33–37, 2006. URL http://dx.doi.org/10.1140/epjcd/s2006-03-004-4.

Ciro Cattuto, Vittorio Loreto, and Vito D.P. Servedio. A yule-simon process with memory. *Europhysics Letters*, 76(2):208–214, 2006.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007b. URL http://xxx.lanl.gov/abs/cs.CY/0605015.

Lukasz Debowski. Zipf's law against the text size: a half-rational model. *Glottometrics*, 4:49 – 60, 2002.

S. N. Dorogovtsev and J. F. F. Mendes. *Phys. Rev. E*, 62:1842, 2000.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968. ISBN 0471257087. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{\&}path=ASIN/0471257087.

Alexander F. Gelbukh and Grigori Sidorov. Zipf and heaps laws' coefficients depend on language. In Alexander F. Gelbukh, editor, *CICLing*, volume 2004 of *Lecture Notes in Computer Science*, pages 332–335. Springer, 2001.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004. ISSN 1046-8188. URL http://portal.acm.org/citation.cfm?id=963770.963772.

Ramon Ferrer i Cancho and Vito D.P.Servedio. Can simple models explain zipf's law for all exponents? *Glottometrics*, 11:1–8, 2005.

Ramon Ferrer i Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *PNAS 788-791*, 100(3):788–791, 2003.

Norman L. Johnson and Samuel Kotz. *Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory*. Wiley, New York, 1977.

J. A. Konstan, J. Reidl, A. Borchers, and J.L. Herlocker. Recommender systems: A grouplens perspective. In *Recommender Systems: Papers from the 1998 Workshop (AAAI Technical Report WS-98-08)*, pages 60–64. AAAI Press, 1998.

András Kornai. How many words are there? *Glottometrics*, 4:61 – 86, 2002.

Wentian Li. Information on zipf's law, 2006. URL http://www.nslij-genetics.org/wli/zipf/index.html.

B. Mandelbrot. An infromational theory of the statistical structure of language. *Communication theory*, 486, 1953.

Benoit Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by h. a. simon. *Information and Control*, 2(1):90–99, April 1959. URL http://dblp.uni-trier.de/db/journals/iandc/iandc2.html#Mandelbrot59.

A.A. Markov. Extension of the law of large numbers to dependent variables. 1951.

Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, 2004. ISSN 1046-8188.

G. A. Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70:311–314, 1957.

Marcelo A. Montemurro and Pedro A. Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10:451, 2002. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0201139.

Marcelo A. Montemurro and D. Zanette. Frequency-rank distribution of words in large text samples: phenomenology and models. *Glottometrics*, 4:87–99, 2002.

Netflix Homepage. Netflix homepage. URL http://www.netflix.com/Default.

Netflix Prize Homepage. Netflix prize homepage. URL http://www.netflixprize.com/.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64 (2):025102, Jul 2001. doi: 10.1103/PhysRevE.64.025102.

M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005. URL doi:10.1080/00107510500052444.

Ben J. Schafer, Joseph A. Konstan, and John Riedi. Recommender systems in e-commerce. In *ACM Conference on Electronic Commerce*, pages 158–166, 1999. URL http://citeseer.ist.psu.edu/benschafer99recommender.html.

I. Schwab, A. Kobsa, and I. Koychev. Learning user interests through positive examples using content analysis and collaborative filtering, 2001. URL citeseer.ist.psu.edu/schwab01learning.html.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425, 1955.

Luc Steels. Semiotic dynamics for embodied agents. *IEEE Intelligent Systems*, 21(3):32–38, 2006. URL http://dblp.uni-trier.de/db/journals/expert/expert21.html#Steels06.

M. Szomszor, C. Cattuto, H. Alani, K. O'Hara, A. Baldassarri, V. Loreto, and V. D. P. Servedio. Folksonomies, the semantic web, and movie recommendation. In *Bridging the Gap between Semantic Web and Web 2.0*, In Proceedings of 4th European Semantic Web Conference (in press), Innsbruck, Austria, 2007.

TAGora. Theoretical tools for modeling and analyzing collaborative social tagging systems. Deliverable D4.1, TAGora project, 2007.

The Internet Movie Database - Wikipedia Entry. The internet movie database - wikipedia entry. URL http://en.wikipedia.org/wiki/Internet_Movie_Database.

The Internet Movie Database (IMDB) Homepage. The internet movie database (IMDB) homepage. URL http://www.imdb.com/.

G. Udny Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Royal Society of London Philosophical Transactions Series B*, 213:21–87, 1925.

Damián H. Zanette and Marcelo A. Montemurro. Dynamics of text generation with realistic zipf's distribution. *Journal of Quantitative Linguistics*, 12(1):29–40, 2005. URL http://dblp.uni-trier.de/db/journals/jql/jql12.html#ZanetteM05.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA (USA), 1949.