Semantic Stability and Implicit Consensus in Social Tagging Streams *

Claudia Wagner
U. of Koblenz & GESIS
claudia.wagner@gesis.org

Philipp Singer
Graz University of Technology
philipp.singer@tugraz.at

Markus Strohmaier
GESIS & U. of Koblenz
markus.strohmaier@gesis.org

Bernardo Huberman
HP labs Palo Alto
bernardo.huberman@hp.com

ABSTRACT

One potential disadvantage of social tagging systems is that due to the lack of a centralized vocabulary, a crowd of users may never manage to reach a consensus on the description of resources (e.g., books, images, users or songs) on the Web. Yet, previous research has provided interesting evidence that the tag distributions of resources in social tagging systems may become semantically stable over time as more and more users tag them and implicitly agree on the relative importance of tags for a resource. At the same time, previous work has raised an array of new questions such as: (i) How can we assess semantic stability in a robust and methodical way? (ii) Does the semantic stabilization varies across different social tagging systems and ultimately, (iii) what are the factors that can explain semantic stabilization in such systems? In this work we tackle these questions by (i) presenting a novel and robust method which overcomes a number of limitations in existing methods, (ii) empirically investigating semantic stabilization in different social tagging systems with distinct domains and properties and (iii) detecting potential causes of stabilization and implicit consensus, specifically imitation behavior, shared background knowledge and intrinsic properties of natural language. Our results show that tagging streams which are generated by a *combination of* imitation dynamics and shared background knowledge exhibit faster and higher semantic stability than tagging streams which are generated via imitation dynamics or natural language phenomena alone.

*A preliminary version of this paper was presented at the World Wide Web Conference (WWW2014) in Seoul, South Korea.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Collaborative computing; H.5.3 [**Group and Organization Interfaces**]: Theory and models

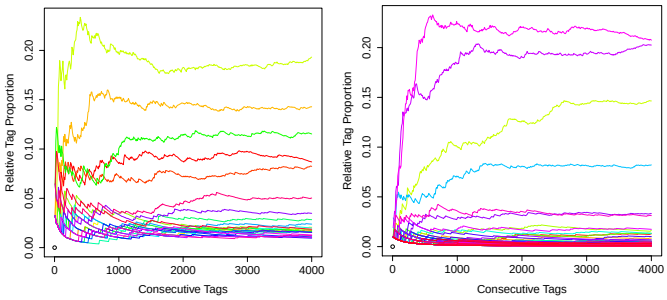
Keywords

social tagging; emergent semantics; social semantics; distributional semantics; stabilization process

1. INTRODUCTION

Instead of enforcing rigid taxonomies or ontologies with controlled vocabulary, social tagging systems allow users to freely choose so-called tags to annotate resources on the Web such as users, books or videos. A potential disadvantage of tagging systems is that due to the lack of a controlled vocabulary which is a central element of traditional forms of organizing information, a crowd of users may never manage to reach a consensus or may never produce a semantically stable description of resources. By semantically stable we mean that users have implicitly agreed on a set of descriptors and their relative importance for a resource which both remain stable over time. That means, *the concept of semantic stability* implies *stability* (i.e., tolerance to deviations or perturbations in time) as well as *implicit consensus* on the description of a resource (i.e., the collective formation of a clear ranking of tags which exposes the relative importance of tags for a resource). Note that implicit consensus and stability are interdependent in social tagging systems. If users do not agree on the description of a resource, they would produce a relatively flat list of descriptors where many descriptors would be equally important for the resource and therefore the ranking of descriptors would be unstable and prone to perturbations.

Yet, when we observe real-world social tagging processes, we can identify interesting dynamics from which a semantically stable set of descriptors may emerge for a given resource. This semantic stability has important implications for the collective usefulness of individual tagging behavior since it suggests that information organization systems can achieve meaningful resource descriptions and interoperability across distributed systems in a decentralized manner [21].



(a) Nathan Fillion (b) Sky Sports

Figure 1: Relative proportion of the top 25 tags (i.e., user list names) assigned to one heavily tagged Twitter user and one moderately tagged Twitter user. The relative tag proportions become stable as more users tag the two sample users. Each line corresponds to one tag.

Semantically stable social tagging streams of resources[1] are not only essential for attaining meaningful resource interoperability across distributed systems and search, but also for learning lightweight semantic models and ontologies from tagging data (see e.g., [27, 29, 23]). Ontologies are often defined as agreed-upon and shared conceptualizations of a domain of interest [12]. Hence, consensus as well as stability are essential for ontologies and can consequently be seen as a prerequisite for learning ontologies from tagging data.

These observations have sparked a series of research efforts focused on (i) methods for assessing semantic stability in tagging streams (see e.g., [10, 13]), (ii) empirical investigations into the semantic stabilization process and the cognitive processes behind tagging (see e.g., [8, 20]) and (iii) models for simulating the tagging process (see e.g., [4, 7]).

Research questions. While previous work makes a promising case for the existence of semantic stabilization in tagging streams, it raises more questions that require further attention, including but not limited to the following: (i) What exactly is semantic stabilization in the context of social tagging streams, and how can we assert it in a robust way? (ii) How suitable are the different methods which have been proposed so far and how do they differ? (iii) Does semantic stabilization vary across different social tagging systems and if yes, in what ways? And finally, (iv) what are the factors that may explain the emergence of semantic stability in social tagging streams?

Contributions. The main contributions of this work are threefold. We start by making a *methodological* contribution. Based on a systematic discussion of existing methods for asserting semantic stability in social tagging systems we identify potentials and limitations. We illustrate these on a previously unexplored people tagging dataset and a synthetic tagging dataset. We explore different subsamples of our dataset including heavily or moderately tagged resources (i.e., a high or moderate amount of users have tagged a resource). Using these insights, we present a novel and flexible method which allows to measure and compare the semantic stabilization in different tagging systems. Flexibility is achieved through the provision of two meaningful parameters, robustness is demonstrated by applying it to random control processes.

Our second contribution is *empirical*. We conduct empirical analysis of semantic stabilization in a series of distinct social tagging systems using our method. We find that the semantic stability of tagging streams in systems which support imitation mechanisms goes clearly beyond what can be explained by the semantic stability of natural language and randomly generated tag distributions drawn from uniform or power law tag distributions. For social tagging systems which do not support imitation we observe the same level of semantic stabilization as for natural language and the synthetic power law tag distributions.

Our final contribution is *explanatory*. We investigate factors which may explain the stabilization and implicit consensus formation processes in social tagging systems. Our results show that tagging streams which are generated by a *combination of* imitation dynamics and shared background knowledge exhibit faster and higher semantic stability than tagging streams which are generated via imitation dynamics or natural language streams alone.

Structure. This paper is structured as follows: We start in Section 2 by highlighting that not all state-of-the-art methods are equally suited for measuring semantic stability in tagging systems, and that some important limitations hinder progress towards a deeper understanding about social-semantic dynamics involved. Based on this discussion, we introduce the data used for our empirical study in Section 3 and present a novel method for assessing semantic stability in Section 4. In Section 5 we aim to shed some light on the factors which may influence the stabilization process. We discuss our results in Section 6 and related work in Section 7. and conclude our work in Section 8.

2. STATE-OF-THE-ART METHODS FOR ASSESSING SEMANTIC STABILIZATION

In the following, we compare and discuss three existing and well-known state-of-the-art methods for measuring stability of tag distributions: *Stable Tag Proportions* [10], *Stable Tag Distributions* [13] and *Power Law Fits* [4]. We define tag distributions of resources as rank-ordered tag frequencies where the frequency of a tag depends on how many users have assigned the tag to a resource. We illustrate the usefulness and limitations of these methods on a previously unexplored people tagging dataset[2] and a synthetic uniformly random tagging dataset which will both be described in Section 3. Each section (i) points out the intuition and definition of the method, (ii) applies the method to the data, and (iii) describes limitations and potentials of the method at hand.

2.1 Method 1: Stable Tag Proportions [10]

Intuition and Definition: In previous work, Golder and Huberman [10] analyzed the relative proportion of tags assigned to a given resource (i.e., $P(t|e)$ where t is a tag and e is an resource) as a function of the number of tag assignments. In their empirical study on Delicious the authors found a stable pattern in which the proportions of tags are nearly fixed for each website after a few hundred tag assignments.

[1]We define a (social) tagging stream as a a temporally ordered sequence of tags produced by a group of users that annotate the same resource.

[2]The limitations of the methods are independent of the dataset and we get similar results using the other datasets introduced in Section 3.

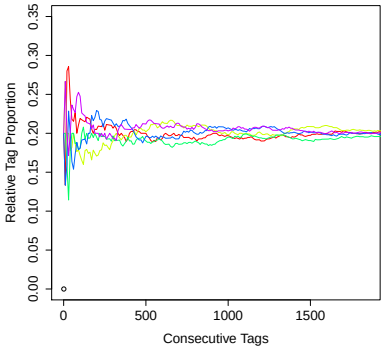


Figure 2: Relative tag proportion of a uniformly random tagging process where each tag assignment on the x-axis corresponds to picking one of the five tags uniformly at random. All tag proportions become relatively stable over time but are all similar. Each line corresponds to one synthetic tag.

Demonstration: In Figure 1 we see that the top tags of a different type of resource (Twitter users rather than websites) also give rise to a stable pattern in which the proportions of tags are nearly fixed. This indicates that, although users keep creating new tags and assign them to resources, the proportions of the tags per resource become stable.

Limitations and Potentials: In [10] the authors suggest that the stability of tag proportions indicates that users have agreed on a certain vocabulary which describes the resource. However, also tag distributions produced by a uniformly random tagging process (see Figure 2) become stable as more tag assignments take place since the growing denominator (i.e., the total sum of the tag frequencies) will flatten any local deviations over time.

However, the stable tagging patterns which are e.g. shown in Figure 1 go beyond what can be explained by a uniformly random tagging process[3] which produces similar proportions for all tags (see Figure 2). Hence, small changes in the tag frequency vector are enough to change the order of the ranked tags (i.e., the relative importance of the tags for the resource). For real tag distributions this is not the case since these tag distributions are distributions with short heads and heavy tails – i.e., few tags are used far more often than most others. We exploit this observation for defining our novel method for assessing semantic stability in Section 4.

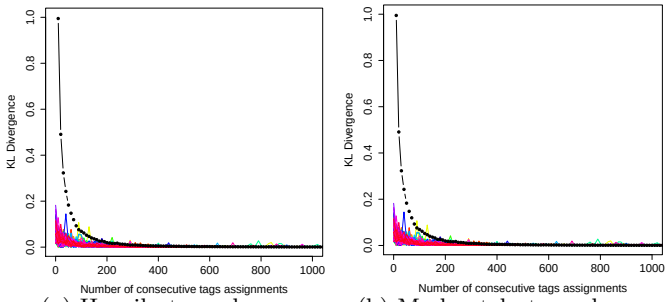
2.2 Method 2: Stable Tag Distributions [13]

Intuition and Definition: Halpin et al. [13] present a method for measuring the semantic stabilization by using the Kullback Leibler (KL) divergence between the tag distributions of a resource at different points in time. The KL divergence between two probability distributions Q and P (where x denotes an element of the distributions) is defined as follows:

$$D_{KL}(P||Q) = \sum_x P(x) ln(\frac{P(x)}{Q(x)}) \quad (1)$$

The authors use the rank-ordered tag frequencies of the 25 highest ranked unique tags per resource at different points in time to compute the KL divergence. They use each month where the tag distribution had changed as a time point instead of using a fixed number of tag assignments as Golder and Huberman [10] did or we do. This is important since their measure, per definition, converge towards zero if the number of tag assignments is constant as shown later.

[3]tags are randomly picked from a synthetic uniform tag distribution



(a) Heavily tagged users (b) Moderately tagged users

Figure 3: KL divergence between the tag distributions at consecutive time points. Each colored line corresponds to one Twitter user, while the black dotted line depicts a randomly simulated tag distribution. One can see that the KL divergence decreases as a function of the number of tag assignments. The KL divergence of a uniformly random tagging process decreases slightly slower than the KL divergence of the real tagging data.

Demonstration: We use the rank-ordered tag frequencies of the 25 highest ranked tags of each resource and a constant number (M) of consecutive tag assignments. We compare the KL divergence of tag distributions after N and $N+M$ consecutive tag assignments. Using a fixed number of consecutive tag assignments allows exploring the properties of a random and uniform tag distribution which is generated by drawing M random samples from a uniform multinomial distribution.

In Figure 3, each point on the x-axis consists of $M = 10$ consecutive tag assignments and N ranges from 0 to 1000. The black dotted line indicates the KL divergence of a uniformly random tag distribution. One can see from this figure that not only the tag distributions of resources (colored lines) seem to converge towards zero over time (with few outliers), but also uniformly random tag distributions (black line) do.

Limitations and Potentials: A single tag assignment in month j has more impact on the shape of the tag distribution of a resource than a single tag added in month $j + 1$, if we assume the number of tags which are added per month is relatively stable over time. However, if the number of tag assignments per resource varies a lot across different months, convergence can be interpreted as semantic stabilization.

This suggests that without knowing the frequencies of tag assignments per month, the measure proposed by Halpin et al. [13] is limited with regard to its usefulness since one never knows whether stabilization can be observed due to the fact that users agreed on a certain set of descriptors and their relative importance for the resource or due to the fact that the tagging frequency in later months was lower than in earlier months. In our work (see Figure 3), we compare the KL divergence of a randomly generated tag distribution with the KL divergence of real tag distributions. This reveals how much faster users reach consensus compared to what one would expect.

Even though we believe this method already improves the original approach suggested by Halpin et al. [13], it is still limited because it requires to limit the analysis to the top k tags. The KL divergence is only defined between two distributions over the same set of tags. We address this limitation with the new method which we propose in Section 4.

2.3 Method 3: Power Law Fits [22]

Intuition and Definition: Tag distributions which follow a power law are sometimes regarded as semantically stable, (i) because of the scale invariance property of power law distributions – i.e., that regardless how large the system grows, the slope of the distribution would stay the same, and (ii) because power law distributions are heavy tail distributions – i.e., few tags are applied very frequently while the majority of tags is hardly used. Adam Mathes [22] originally hypothesized that tag distributions in social tagging systems follow a power law function. Several studies empirically show that the tag distributions of resources in social tagging systems indeed follow a power law [28, 17, 3, 4]. A power law distribution is defined by the function:

$$y = cx^{-\alpha} + \epsilon \tag{2}$$

Both c and α are constants characterizing the power law distribution and ϵ represents the uncertainty in the observed values. The most important parameter is the scaling parameter α as it represents the slope of the distribution [2, 5]. It is also important to remark that real world data nearly never follows a power law for the whole range of values. Hence, it is necessary to find some minimum value $xmin$ for which one can say that the tail of the distribution[4] with $x \geq xmin$ follows a power law [5].

Demonstration: We first visualize the rank frequency tag distributions (see Figure 4(a) and Figure 4(b)) and the complementary cumulated distribution function (CCDF) of the probability tag distributions (see Figure 4(c) and Figure 4(d)) on a log-log scale. We see that for heavily and moderately tagged resources, few tags are applied very frequently while the vast majority of tags are used very rarely. Figure 4(c) and Figure 4(d) show that the tag distributions of heavily and moderately tagged resources are dominated by a large number of tags which are only used once.

Figure 4 reveals that the tails of the tag distributions (starting from a tag frequency 2) are close to a straight line. The straight line, which is a main characteristic for power law distributions plotted on a log-log scale, is more visible for heavily tagged resources than for moderately tagged ones. We can now hypothesize that a power law distribution could be a good fit for our data if we look at the tail of the distribution with a potential $xmin \geq 2$.

For finding the scaling parameter α we use a *maximum likelihood estimation* and for finding the appropriate $xmin$ value we use the *Kolmogorov-Smirnov statistic* as suggested by Clauset et al. [5]. As proposed in previous work [2, 5], we also look at the Kolmogorov-Smirnov distance D of the corresponding fits – the smaller D the better the fit. Table 1 shows the parameters of the best power law fits, averaged over all heavily tagged or moderately tagged resources. One can see from this table that the α values are very similar for both datasets and also fall in the typical range of power law distributions. Further, one can see that the power law fits are slightly better for heavily tagged resources than for moderately tagged ones, as also suggested by Figure 4.

Although our results suggest that it is likely that our distributions have been produced by a power law function, further investigations are warranted to explore whether other heavy-tailed candidate distributions are better fits than the power law [5, 1]. We compare our power law fit to the fit of the *exponential function*, the *lognormal function* and the *stretched exponential (Weibull) function*. We use *log-likelihood ratios* to indicate which fit is better.

The exponential function represents the absolute minimal candidate function to describe a heavy-tailed distribution. That means, if the power law function is not a better fit than the exponential function, it is difficult to assess whether the distribution is heavy-tailed at all. The lognormal and stretched exponential function represent more sensible heavy-tailed functions. Clauset et al. [5] point out that there are only a few domains where the power law function is a better fit than the lognormal or the stretched exponential.

Our results confirm this as we do not find significant differences between the power law fit and the lognormal fit (for both heavily and moderately tagged users). However, most of the time the power law function is significantly better than the stretched exponential function and the power law function is a significantly better fit than the exponential function for all heavily tagged users and for most moderately tagged users. This indicates that the tag distributions of heavily tagged resources and most moderately tagged resources are clearly heavy tail distributions and the power law function is a reasonable well explanation. Nonetheless, it remains unclear from which heavy tail distribution the data has been drawn since several of them produce good fits.

Limitations and Potentials: As we have shown, one limitation of this method is that it is often difficult to determine which distribution has generated the data since several distributions with similar characteristics may produce an equally good fit. Furthermore, the automatic calculation of the best $xmin$ value for the power law fit has certain consequences since $xmin$ might become very large and therefore the tail to which the power law function is fitted may become very short. Finally, there is still an ongoing discussion about the informativeness of scaling laws (see [16] for a good overview), since some previous work suggests that there exist many ways to produce scaling laws and some of those ways are idiosyncratic and artifactual [26, 18].

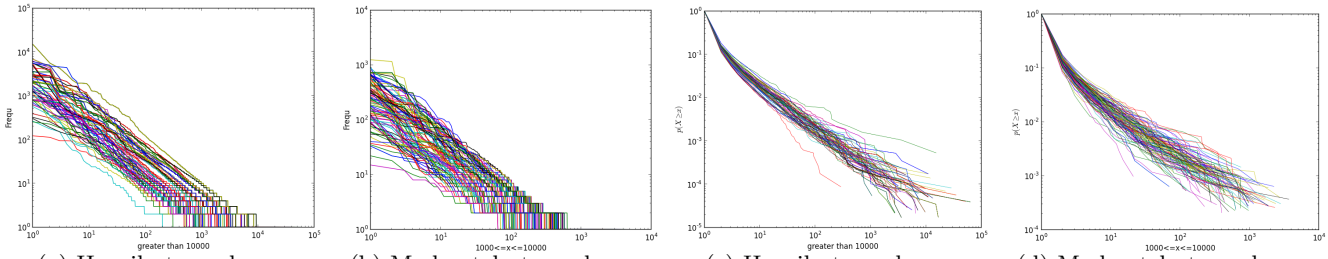
3. EXPERIMENTAL SETUP AND DATASETS

We empirically analyze the semantic stabilization process in a series of different social tagging systems using the state-of-the-art methods described in Section 2 and using a new method introduced in Section 4. Table 2 gives an overview

[4]We use the term *tail* to characterize the end of a distribution in the sense of probability theory.

Table 1: Parameters of the best power law fits.

	α	std	$xmin$	std	D	std
Heavily tagged users	1.9793	0.0841	4.5500	1.9818	0.0299	0.0118
Moderately tagged users	2.0558	0.1529	3.1200	0.0570	0.0570	0.0218



(a) Heavily tagged users (b) Moderately tagged users (c) Heavily tagged users (d) Moderately tagged users

Figure 4: Rank-ordered tag frequency and CCDF plots for heavily tagged and moderately tagged users on log-log scale. The illustrations show that for both heavily and moderately tagged resources, few tags are applied very frequently while the vast majority of tags is applied very rarely. In Figure 4(c) and Figure 4(d) we can see that a large number of tags are only used once. The figures visualizes that the tails of the tag distributions are close to a straight line which suggests that the distributions might follow a power law.

of the datasets obtained from distinct tagging systems using the nature of the resource being tagged, the sequential order of the tagging process (i.e., is the resource selected first or the tag), the existence or absence of tag suggestions and the visibility of the tags which have been previously assigned to a resource. We say that tags have a low visibility if users do not see them during the tagging process and if they are not shown on the page of the resource being tagged. Otherwise, tags have a high visibility. Also, the number of resources, users and tags per dataset are shown.

Delicious dataset: Delicious is a social tagging system where users can tag any type of website. We use the Delicious dataset crawled by Görlitz et al. [11]. From this dataset we randomly selected 100 websites which were tagged by many users (more than 4k users) and 100 websites which were moderately tagged (i.e., by less than 4k but more than 1k users) and explore the consecutive tag assignments for each website. The original dataset is available online[5].

LibraryThing dataset: LibraryThing is a social tagging system which allows to tag books. We use the LibraryThing dataset which was crawled by Zubiaga et al. [34]. Again, we randomly sampled 100 books that were heavily tagged (more than 2k users) and 100 books which were moderately tagged (less than 2k and more than 1k users) and explore the consecutive tag assignments for each book.

Twitter dataset: Twitter is a microblogging service that allows users to tag their contacts by grouping them into user lists with a descriptive title. The creation of such list titles can be understood as a form of tagging since list titles are free form words which are associated with one or several resources (in this case users). What is unique about this form of tagging is that the tag (aka the list title) is usually produced first, and then users are added to this list, whereas in more traditional tagging systems such as Delicious, the process is the other way around. That means, the user first creates a new tag (see Figure 5) and second looks for users he wants to assign to this tag (see Figure 6). Users are not provided with any tag (aka the list title) suggestions. If they want to see which other tags have previously been assigned to their contacts they need to visit the profile page of each users and navigate to their list membership section. Since this is fairly time intensive we can speculate that it is unlikely that users imitate the previously assigned tags but create their own tags and assign users to them based on what they know about them and how they want to organize them.

From a Twitter dataset which we described in previous work [30], we selected a sample of 100 heavily tagged users

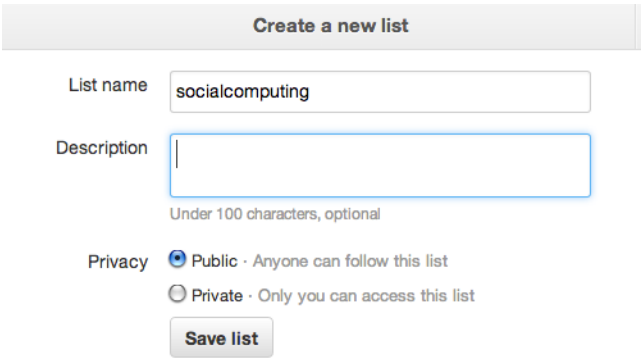


Figure 5: User creates a new list on Twitter.

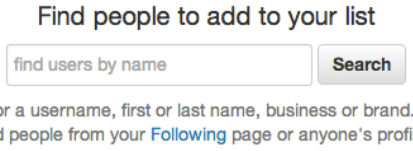


Figure 6: User can assign users to his newly created list by searching people on Twitter or by browsing through the list of users he/she is following.

[5] http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/DataSets/PINTSExperimentsDataSets

Table 2: Description of the datasets and characteristics of the social tagging system the data stem from.

System	Entity Type	Tag First	Tag Suggestions	Tags Visible	#Resources	#Users	#Tags
Delicious	websites	no	yes	low	17,000k	532k	2,400k
LibraryThing	books	no	no	high	3,500k	150k	2,000k
Twitter lists	users	yes	no	low	3,286	2,290k	1,111k

(which are mentioned in more than 10k lists) and 100 moderately tagged users (which are mentioned in less than 10k lists and more than 1k lists). For each of these sample users we crawled the full history of lists to which a user was assigned. We do not know the exact time when a user was assigned to a list but we know the relative order in which a user was assigned to different lists. Therefore, we can study the tagging process over time by using consecutive list assignments as a sequential ordering[6].

It needs to be noted that the thresholds we have used above during the data collection are distinct for each tagging system since those systems differ amongst others in their number of active users and size. We chose the thresholds empirically and found that the choice of threshold does not impact our results since heavily tagged as well as moderately tagged resources show similar characteristics.

Finally, we also contrast our tagging datasets with a natural language corpus (see Section 5.2) and two randomly generated tagging dataset. This allows us on one hand, to explore to what extent semantic stabilization which can be observed in tagging systems goes beyond what one would expect to observe if the tagging process would be a random process; and on the other hand, to compare the semantic stabilization of the tag distributions of resources with the semantic stabilization of co-occurring word distributions of resources.

Natural Language corpus: As a natural language corpus we use a sample of tweets which refer to the same resource. Therefore, we selected a random sample of users from our Twitter dataset which have received tweets from many distinct users (more than 1k). For those users, we select a sample of up to 10k tweets they received. The words of those tweets are extracted and interpreted as social annotations of the receiver. This allows us to compare tags with words, both annotating a resource (in this case a user). We removed URLs, usernames, punctuations, numbers and Twitter syntax such as RT using the part of speech tagger presented in [9].

Synthetic random tagging datasets: Given a fixed vocabulary size we create two random tagging dataset by simulating the tagging process as random draws from a uniform and a power law tag distribution.

4. MEASURING SEMANTIC STABILITY

Based on the analysis of state-of-the-art methods presented in Section 2, we (i) present a novel method for assessing the semantic stability of individual tagging streams and (ii) show how this method can be used to assess and compare the stabilization process in different tagging systems. Our new method incorporates three new ideas:

Ranking of tags: A tagging stream can be considered as semantically stable if users have implicitly agreed on a ranking of tags which remains stable over time. Importantly, the ranking of frequent tags remains more stable than the ranking of less frequent tags since frequent tags are those which might be more relevant for a resource. They have been applied by many users to a resource and therefore stable rankings of these tags indicate that a large group of users has agreed on the relative importance of the tags for that resource.

Random baselines: Semantic stability of random tagging processes needs to be considered as a baseline for stability since we are interested in exploring stable patterns which go beyond what can be explained by a random tagging process.

New tags over time: New tags can be added over time and therefore, a method which compares the tag distributions of one resource at different points in time must be able to handle mutually non-conjoint tag distributions – i.e., distributions which contain tags that turn up in one distribution but not in the other one. Most measures used in previous work (e.g., the KL divergence) only allow to compare the agreement between mutually conjoint lists of elements and a common practice is to prune tag distributions to their top k elements – i.e., the most frequently used tags per resource. However, this pruning requires global knowledge about the tag usage and only enables a post-hoc rather than a real-time analysis of semantic stability.

4.1 Rank Biased Overlap: $RBO(\sigma 1, \sigma 2, p)$

Intuition and Definition: The Rank Biased Overlap (RBO) [31] measures the similarity between two rankings and is based on the cumulative set overlap. The set overlap at each rank is weighted by a geometric sequence, providing both top-weightedness and convergence. RBO is defined as follows:

$$RBO(\sigma 1, \sigma 2, p) = (1-p)\sum_{d=1}^{\infty} \frac{\sigma 1_{1:d} \cap \sigma 2_{1:d}}{d} p^{(d-1)} \qquad (3)$$

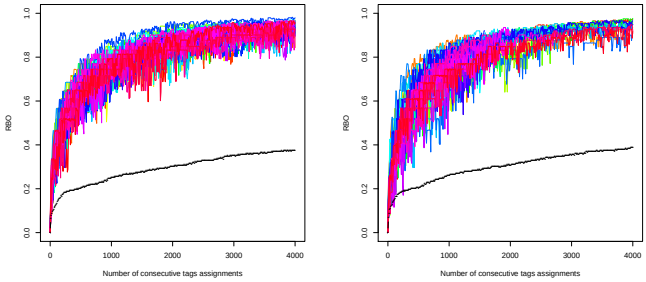
Let $\sigma 1$ and $\sigma 2$ be two not necessarily conjoint lists of ranking. Let $\sigma 1_{1:d}$ and $\sigma 2_{1:d}$ be the ranked lists at depth d. The RBO falls in the range $[0, 1]$, where 0 means disjoint, and 1 means identical. The parameter p ($0 \leq p < 1$) determines how steep the decline in weights is. The smaller p is, the more top-weighted the metric is. If $p = 0$, only the top-ranked item of each list is considered and the RBO score is either zero or one. On the other hand, as p approaches arbitrarily close to 1, the weights become arbitrarily flat. These weights, however, are not the same as the weights that the elements at different ranks d themselves take, since these elements contribute to multiple agreements.

In the following, we use a version of RBO that accounts for tied ranks. As suggested in [31], ties are handled by assuming that if t items are tied for ranks d to $d+(t-1)$, they all occur at rank d. RBO may account for ties by dividing twice the overlap at depth d by the number of items which occur at depth d, rather than the depth itself:

$$RBO(\sigma 1, \sigma 2, p) = (1-p)\sum_{d=1}^{\infty} \frac{2 * \sigma 1_{1:d} \cap \sigma 2_{1:d}}{|\sigma 1_{1:d} + \sigma 2_{1:d}|} p^{(d-1)} \qquad (4)$$

We modify RBO by summing only over occurring depths rather than all possible depths. Therefore, our RBO measure further penalizes ties and assigns a lower RBO value to pairs of lists containing ties. For example, consider the following two pairs of ranked lists of items: (i) *(A=1, B=2, C=3, D=4)*, *(A=3, B=2, C=1, D=4)* and (ii) *(A=1, B=1, C=1, D=4)*, *(A=1, B=1, C=1, D=4)*. Both pairs of lists have the same concordant pairs: (A,D) and (B,D) and (C,D). The RBO value of the first pair is 0.2 according to the origi-

[6]We share the Twitter user handles to allow other researchers to recreate our dataset and reproduce our results for our heavily tagged `http://claudiawagner.info/data/gr_10k_username.csv` and moderately tagged `http://claudiawagner.info/data/less_10k_username.csv` Twitter users.



(a) Heavily tagged users (b) Moderately tagged users

Figure 7: Rank Biased Overlap (RBO) measures with $p = 0.9$. The black dotted line shows the weighted average RBO of a uniformly random tagging process over time, while each colored line corresponds to the RBO of one Twitter user.

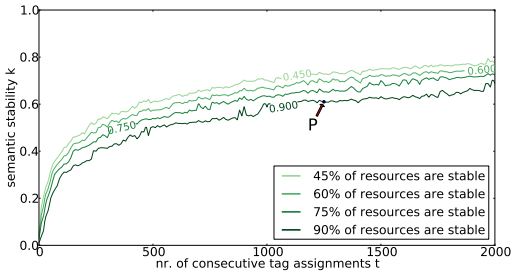


Figure 8: The percentage of resources (in this case heavily tagged Twitter users) stabilized at time t with stability threshold k. For example, point P indicates that after 1250 tag assignments 90% of resources exhibit semantic stability (an RBO value) of 0.61 or higher.

nal measure and also according to our tie-corrected variant. The RBO value of the second pair is 0.34 according to the original measure and 0.17 according to our tie-corrected variant. This example nicely shows that while the original RBO measure tends to overestimate ties, our variant slightly penalizes ties. For our use case this makes sense since we do not want to overestimate the semantic stability of a resource where users have not agreed on a ranking of tags but only find that all of tags are equally important.

Demonstration: Figure 7 shows the RBO of the tag distributions of resources over time for our people tagging dataset. The RBO value between the tag distribution after N and $N + M$ tag assignments is high if the M new tag assignments do not change the ranking of the (top-weighted) tags. One can see that the RBO of a randomly generated tag distribution is pretty low and increases slowly as more and more tags are added over time. Contrary, the RBO of real tag distributions increases as more and more tags are added. At the beginning, it increases quickly and remains relatively stable after few thousand tag assignments. This indicates that RBO allows identifying an implicit consensus in the tag distributions which may emerge over time and which *goes beyond what one would expect from a uniformly random tagging process*. A uniformly random tagging process produces relative tag proportions which are all very similar (i.e., all tags are equally important or unimportant). Therefore, the probability that the ranking changes after new tag assignments is higher than it is for real tagging streams where users have produced a clear ranking of tags where some tags are much more important for a resource than others. Over time, the gap between real tagging streams and random tagging streams will decrease. Yet, one can see that within the time-window in which real tagging streams semantically stabilize (i.e., few thousand tag assignments) tag distributions produced by a random process are significantly less stable. Again, we can see that the tag distributions of heavily tagged resources are slightly more stable than those of moderately tagged ones.

In our work, we empirically chose $p = 0.9$ which means that the first 10 ranks have 86% of the weight of the evaluation. We got similar results when choosing higher values of p. For example, when choosing $p = 0.98$ the first 50 items get 86% of the weight. If one would chose a lower value for p such as $p = 0.1$ (or $p = 0.5$) the first two elements would get 99.6% (or 88.6%) of the weight. That means, all elements with a rank lower than two would be almost ignored and therefore the RBO values show more fluctuation. However, in all our experiments with different p values the RBO of real tag distributions was significantly higher than the RBO of uniform random tag distributions.

Limitations and Potentials: One advantage of RBO is that it handles mutually non-conjoint lists of tags, weights highly ranked tags more heavily than lower ranked tags, and is monotonic with increasing depth of evaluation. Another advantage is that rank agreements are measured rather than distances between distributions. While the deviations in the distributions decrease with increasing denominators, the deviations in the rankings do not necessarily decrease. The probability of observing local deviations in the rankings depends on the shape of the distribution, the flatter the distribution the more likely the rankings will continue changing over time.

A potential limitation of RBO is that it requires to pick the parameter p which defines the decline in weights - i.e., how top-weighted the RBO measure is. Which level of top-weightness is appropriate for the tag distributions in different tagging systems might be a controversial question. However, our experiments revealed that as long as the parameter p was not chosen to be small (i.e., $p < 0.5$), the results remained essentially the same.

4.2 A Rank-based Semantic Stability Method

Based on the previously defined *Rank Biased Overlap* we propose a method which allows to investigate the semantic stabilization process in a social tagging system (or other systems in which social streams are generated) based on the stabilization and consensus formation process of individual social tagging streams of resources. This method allows to compare the semantic stabilization process of different social stream based systems over time. Given a sample of tagged resources (the sample size N and the type of resources can be chosen arbitrarily) the goal is to specify how many resources of the sample have stabilized after a certain number of consecutive tag assignments.

We propose a flexible and fluid definition of the concept of *semantic stabilization* by introducing (a) a parameter k that constitutes a threshold for the RBO value and (b) a parameter t that specifies the number of consecutive tag assignments. We call a resource in a social tagging system *semantically stable at point t*, if the RBO value between its tag distribution at point $t-1$ and t is equal or greater than

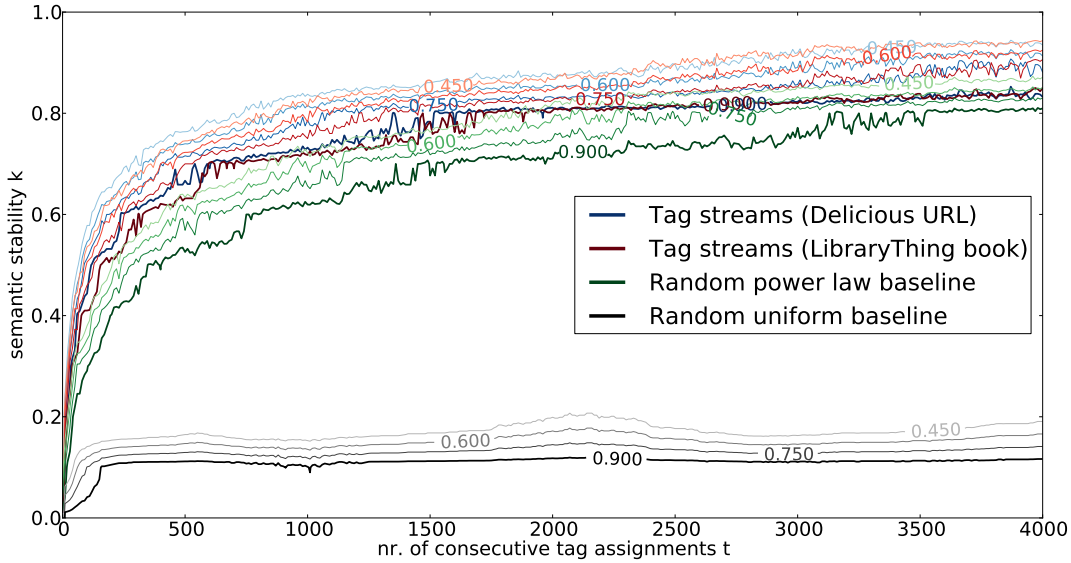


Figure 9: Semantic stabilization in different social tagging datasets which goes beyond what one would expect from two random control processes. The x axis represents the consecutive tag assignments t while the y-axis depicts the RBO (with $p = 0.9$) threshold k. The contour lines illustrate the curve for which the function $f(t, k)$ has constant values. These values are depicted in the lines and represent the percentage of stabilization f. Each dataset is represented by a distinct color map. Lines which belong to the same color map show for each number of tag assignments t the k threshold for which 90%, 75%, 60% and 45% of all resources have an RBO value equal or higher than k. On can see that tagging streams in Delicious and LibraryThing stabilize faster and reach higher levels of semantic stability than one would expect according to both baseline tagging processes.

k. Our proposed method allows to calculate the percentage of resources that have semantically stabilized after a number of consecutive tag assignments t according to some threshold for stabilization k. We can define this function by:

$$f(t,k) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & \text{if } RBO(\sigma_{i_{t-1}}, \sigma_{i_t}, p) > k.\\ 0, & \text{otherwise.}\end{cases} \quad (5)$$

We illustrate the semantic stabilization for our sample of heavily tagged Twitter users in Figure 8. The contour plot depicts the percentage of resources (i.e., Twitter users) which have become semantically stable according to some RBO threshold k after t tag assignments. The figure shows that after 1k tag assignments 90% of Twitter users have an RBO value above 0.5 which can be considered as a medium level of stability. We define RBO values below 0.4 as a sign for no stability, values between 0.4 and 0.7 as medium stability and values above 0.7 as high stability.

4.2.1 Results & Discussion

In this section we use our novel method to explore and compare the semantic stabilization process of different social tagging systems introduced in Section 3.

The contour plots in Figure 9 and Figure 10 depict the percentage of resources which have become semantically stable (i.e., users have agreed on a stable and focused list of tags) according to some RBO threshold k after t tag assignments in different social tagging systems. Two randomly generated tagging datasets are added for control. One is drawn from a uniform tag distribution and the other one is drawn from a tag distribution which follows a power law. Figure 9 only includes social tagging systems which semantically stabilize faster and reach higher levels of semantic stability than both baselines, while Figure 10 only includes those social tagging systems which do not beat both baselines.

In both figures we can see that the tag distributions generated by a uniform tagging process[7] exhibits by far the lowest stabilization since the resources just stabilize for low k ($k < 0.2$) even after a large amount of tag assignments t. That means, the k threshold for which 90%, 75%, 60% and 45% of all resources have an equal or higher RBO values than k is very low. Contrary, we can see that all real-world tagging systems exhibit much higher stability and consensus. The tag distributions that are generated by a random power law tagging process[8] show higher and faster stabilization. However, Figure 9 shows that some real-world tagging systems such as Delicious and LibraryThing reveal higher and faster semantic stabilization than both baselines – i.e., they show higher k values for lower t values. It is interesting to note that both of these systems encourage imitation behavior by suggesting previously assigned tags (see Delicious) and by making previously assigned tags visible during the tagging process (see LibraryThing).

In Twitter users first have to create a tag (aka user list) and afterwards select the resources (aka users) to which they want to assign the tag. During this tagging process, tags which have been previously assigned to users are not visible and therefore it is unlikely that imitation behavior plays a major role in Twitter[9].

[7] tags are randomly picked from a synthetic uniform tag distribution

[8] tags are randomly picked from a synthetic tag distribution that follows a power law

[9] If users want to see which other tags have previously been assigned to a user they need to visit her profile page and navigate to the list membership section. Since this is fairly

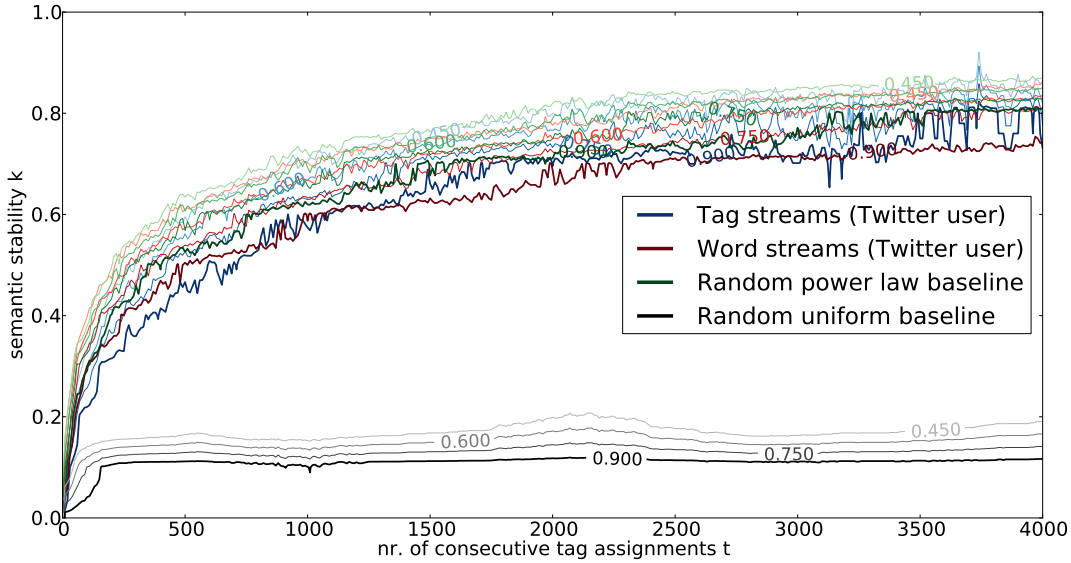


Figure 10: Semantic stabilization in different social tagging datasets and a natural language dataset. One can see that there is no significant difference between the stabilization and consensus formation process in the synthetically generated tag streams which have been drawn from a power law distribution, the natural language and the tag streams in Twitter. The x axis represents the consecutive tag assignments t while the y-axis depicts the RBO (with $p = 0.9$) threshold k. The contour lines illustrate the curve for which the function $f(t, k)$ has constant values. These values are depicted in the lines and represent the percentage of stabilization f.

Interestingly, our results show (cf. Figure **??**) that there are no significant differences between the stabilization and consensus formation process of synthetically generated tag streams that have been drawn from a power law distribution, natural language streams and the tag streams in Twitter. All three streams reach a medium level of semantic stability for different values of t which is significantly lower than the semantic stability level reached by social tagging systems which support imitation. Our empirical results from different social tagging systems are in line with the results from a user study presented in [2] which also shows that tag distributions of resources become stable regardless of the visibility of previously assigned tags. The presence of tag suggestions may provoke a higher and faster agreement between users who tag a resource and may therefore lead to higher levels of stability, but it is clearly not the only factor causing stability. Our results suggest that in tagging systems which encourage imitation less than 1k tag assignments are necessary before a tagging stream becomes semantically stable (i.e., the rank agreement has reached a certain level and does not change anymore), while in tagging systems which do not encourage imitation more than 1k tag assignments are required.

5. EXPLAINING SEMANTIC STABILITY

The experimental results reported in [2] as well as our own empirical results on the people tagging dataset from Twitter suggest that stable patterns may also arise in the absence of imitation behavior. As a consequence, other factors that might explain semantic stabilization, such as shared background knowledge and stable properties of natural language, deserve further investigation.

5.1 Imitation and Background Knowledge

To explore the potential impact of imitation and shared background knowledge we simulate the tag choice process. According to [7] there are several plausible ways how the tag choice process can be modeled:

Random tag choice: Each tag is chosen with the same probability. This corresponds to users who randomly choose tags from the set of all available tags which seems to be only a plausible strategy for spammers

Imitation: The tags are chosen with a probability that is proportional to the tag's occurrence probability in the previous stream. This selection strategy corresponds to the Polya Urn model described in [10] where only tags that have been used before are in the urn and can be selected. Users who are easily influenced by other users might apply this tag selection strategy.

Background Knowledge: The tags are chosen with a probability that is proportional to the tag's probability in the shared background knowledge of users. This corresponds to users who choose tags that seem appropriate based on their own background knowledge.

In our simulation, we assume that the tag choice of users might be driven by both imitation and background knowledge. Similar to the epistemic model [7] we introduce a parameter I describing the impact of imitation. Consequently, the impact of shared background knowledge is $1-I$. We run I from 0 to 1 – i.e., we simulate tagging streams which have been generated by users who only use the imitation strategy to choose their tags ($I = 1$), users who only rely on their background knowledge when selecting tags ($I = 0$), and users who adapt both strategies. We use

time intensive one can speculate that it is unlikely that users imitate the previously assigned tags but create their own tags and assign users to them based on what they know about them and how they want to organize them.

a word-frequency corpus[10] from Wikipedia to simulate the shared background knowledge. That means, if only background knowledge ($I = 0$) is used, we sample tags from the word frequency distribution of Wikipedia. Since this distribution follows Zipf law, we do not need to include our second baseline for which we synthetically generate tag distributions by drawing them from a power law distribution because they would be identical. If only imitation is used ($I = 1$) the first user picks a tag which is added to the urn and afterwards the Polya Urn model [10] is used. That means, in an extreme tagging scenario where every user always imitates the previous user, the same tag will be re-assigned to the resource. For each synthetic dataset we simulate 100 tagging streams in order to have the same sample size as for our real-world datasets introduced in Section 3.

Our results in Figure 11 show the percentage of resources which have a RBO value equal or higher than k after t tag assignments for different synthetic tagging datasets. One can see from this figure that a synthetic tagging dataset with $I = 1$ (i.e., a datasets which was solely created via imitation behavior) does not stabilize over time since more than 90% of the resources have very low RBO values (i.e., $k < 0.1$) also after a few thousand tag assignments. This is consistent with our intuition since a model which is purely based on imitation dynamics fails to introduce new tags and therefore no ranked lists of tags per resource can be created.

Further, one can see that a synthetic tagging dataset with $I = 0$ (i.e., a tagging datasets which was solely created via background knowledge and therefore reflects the properties of a natural language system) stabilizes slightly slower than a synthetic tagging dataset which was generated by a mixture of background knowledge and imitation dynamics ($I = 0.7$). This is particularly interesting since it suggests that *when shared background knowledge (encoded in natural language) is combined with social imitation, tagging streams reach higher levels of semantic stability (*$0.7 < k < 0.8$*) quicker (for lower* t*) than if users either only rely on imitation behavior or on background knowledge.* Our findings are in line with previous research [7] which showed that an imitation rate between 60% and 90% is best for simulating real tag streams of resources. However, unlike our work their work focuses on reproducing the sharp drop between rank 7 and 10 in the rank-ordered tag frequency distribution of a resource at one time point and does not explore the stabilization process over time. However, as described in Section 7 their work has certain limitations which we address by (i) exploring a range of different social tagging systems including one where no tags are suggested and previously assigned tags are not visible during the tagging process and (ii) studying the semantic stabilization process over time rather than the shape of the rank-ordered tag frequency distribution at a single time point.

5.2 Stability and Consensus in Natural Language Streams

Since tagging systems are natural language systems, the regularities and the stability of natural language (see e.g., [33] and [15]) may cause the stable patterns which we observe in tagging systems. That means, one can argue that tagging systems become stable because they are built on top of natural language which itself is stable.

[10] `http://www.monlp.com/2012/04/16/calculating-word-and-n-gram-statistics-from-a-wikipedia-corpora/`

Our results presented in Figure 10 show that a natural language corpus (see Section 3) – where users talk about a set of sample resources – also becomes semantically stable and reaches a medium level of semantic stability (with $k > 0.6$ if $t > 1,000$), but does not go beyond what one would expect from a random process which draws tags a power law tag distribution. According to the Wilcoxon rank sum test with continuity correction the differences between the stabilization levels of the natural language streams, the tagging streams in Twitter and the random power law tag distributions after different numbers of tag assignments t are not significant.

Also, our simulation results in Figure 11 show that a synthetic dataset which is generated using Wikipedia word frequencies as background knowledge ($I = 0$) and is therefore reflecting the properties of the natural language, becomes semantically stable over time and reaches a medium level of stability and consensus (with $k > 0.6$ if $t > 1,000$). In both cases one can see that the stabilization and consensus formation process of natural language systems clearly differs from the stabilization process of real tagging streams which are produced in systems supporting imitation and synthetic tagging streams which are generated by included imitation mechanisms. The RBO curve of natural language systems is flatter at the beginning than the RBO curve of tagging streams which are partly generated via imitation mechanisms which suggests that more word assignments are needed until a high percentage of resources have RBO values at or above a certain threshold k. The only tagging stream dataset which shows a similar stabilization process as the natural language dataset is the people tagging dataset obtained from Twitter which does not support any imitation mechanisms. This suggests, that the stability of natural language systems can indeed explain a large proportion of the stability which can be observed in tagging systems where the tagging process is not really social (i.e., each user annotates a resource separately without seeing the tags others used) and no imitation dynamics are supported. However, tagging systems which support the social aspect of tagging by e.g., showing tags which have been previously applied by others, exhibit a faster and higher level of semantic stabilization than tagging systems which do not implement these social functionalities. This suggests that the semantic stability which can be observed in *social* tagging systems goes beyond what one would expect from natural language systems and that higher and faster degree of stability and consensus are achieved through the social dynamics in tagging systems; concretely, the imitation behavior of users.

6. DISCUSSION

The main implications of our work are: (i) We highlight limitations of existing methods for measuring semantic stability in social tagging streams and introduce a new and more robust method which allows to analyze the stabilization and consensus formation process in social tagging systems. However, our method is not limited to social tagging systems and tagging streams and can be used to measure stability and user agreement in other types of data streams, such as word-streams of hashtags in Twitter or word streams of Wikipedia concepts. (ii) Our empirical results as well as our simulation results suggest that when aiming to improve semantic stability of social tagging systems, system designers can exploit the insights gained from our work by

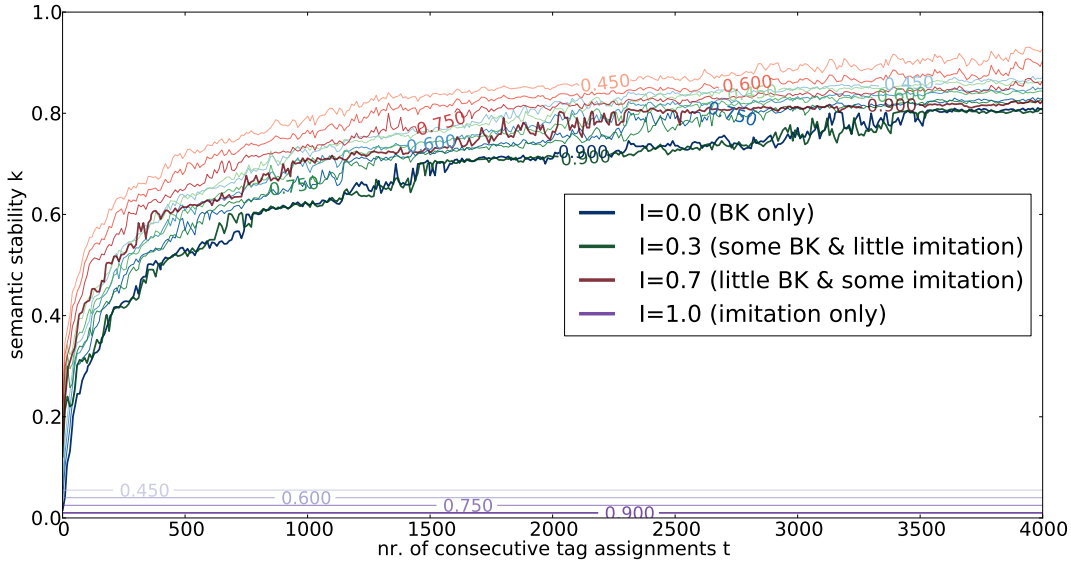


Figure 11: Semantic stabilization of synthetic (i.e., simulated) tagging processes. Tagging streams which are generated by a combination of imitation dynamics (70%) and background knowledge (30%) tend to stabilize faster and reach higher levels of stability than streams which are generated by imitation behavior (I=1) or background knowledge (I=0) alone.

implementing mechanisms which - for example - augment imitation in 70% of cases (e.g., by suggesting or showing previously assigned tags) while tapping into the background knowledge of users in 30% of cases (e.g., by requiring users to tag without recommendation mechanisms at place, thereby utilizing background knowledge).

In future we also want to explore the lowest number of users that need to tag a resource in order to produce a stable tag description of the resource for which we would also need to model the number of tags users simultaneously assign to resources into our experiments. Further, we want to point out that for the sake of simplicity we used the same background knowledge corpus for all resources and neglected the impact of the user interface (i.e., the number of suggested tags and the number of previously used tags from which they are chosen) on the imitation process. These user interface parameters are different for distinct tagging systems and have been varied over time. Without exactly knowing how the user interface looked like when the data was generated and how the algorithm for suggesting and displaying tags worked, it is difficult to properly choose these parameters.

7. RELATED WORK

Social tagging systems have emerged as an alternative to traditional forms of organizing information which usually enforce rigid taxonomies or ontologies with controlled vocabulary. Social tagging systems, however, allow users to freely choose so-called tags to annotate resources such as websites, users, books, videos or artists.

In past research, it has been suggested that stable patterns may emerge when a large group of users annotates resources on the Web. That means, users seem to reach a consensus about the description of a resource over time, despite the lack of a centralized vocabulary which is a central element of traditional forms of organizing information [10, 13, 4]. Several methods have been established to measure this semantic stability: (i) in previous work one co-author of this paper suggested to assess semantic stability by analyzing the proportions of tags for a given resource as a function of the number of tag assignments [10]. (ii) Halpin et al. [13] proposed a direct method for quantifying stabilization by using the Kullback-Leibler (KL) divergence between the rank-ordered tag frequency distributions of a resource at different points in time. (iii) Cattuto et al. [4] showed that power law distributions emerge when looking at rank-ordered tag frequency distributions of a resource which is an indicator of semantic stabilization. Lin et al. [19] investigate dynamic properties of social tagging systems (e.g., tag growth) on a macro level (i.e., per system) and on a micro level (i.e., per resource). They analyze amongst others the tag growth in different systems and argue that a slower growth rate indicates a larger portion of tag reusing, which further implies a stronger collective feedback and higher level consensus over time. The semantic stability measure which we propose in our work goes beyond the notion of stability and consensus described by Lin et al. since the relative importance of tags for a resource may change over time though no new tags are introduced. Therefore, we believe that a slow tag growth rate is not a sufficient criteria for semantic stability, though it might often be observed in social tagging systems where users imitate the tags of others.

Several attempts and hypotheses aiming to explain the observed stability have emerged. In [10] the authors propose that the simplest model that results in a power law distribution of tags would be the classic Polya Urn model. The first model that formalized the notion of new tags was proposed by Cattuto et al. [4] by utilizing the Yule-Simon model [32]. Also, models like the semantic imitation model [8] or simple imitation mechanisms [20] have been deployed for explaining and reconstructing real world semantic stabilization.

While above models mainly focus on the imitation behavior of users for explaining the stabilization process, shared background knowledge might also be a major factor as one co-author of this work already hypothesized in previous work [10]. Research by Dellschaft et al. [7] picked up this hypoth-

esis and explored the utility of background knowledge as an additional explanatory factor which may help to simulate the tagging process. Dellschaft et al. show that combining background knowledge with imitation mechanisms improves the simulation results. Although their results are very strong, their evaluation has certain limitations since they focus on reproducing the sharp drop of the rank-ordered tag frequency distribution between rank 7 and 10 which was previously interpreted as one of the main characteristics of tagging data [3]. However, recent work by Bollen et al. [2] questions that the flatten head of these distributions is a characteristic which can be attributed to the tagging process itself. Instead, it may only be an artifact of the user interface which suggests up to ten tags. Bollen et al. show that power law forms regardless of whether tag suggestions are provided to the user or not, making a strong point towards the utility of background knowledge for explaining the stabilization.

In addition to imitation and background knowledge, an alternative and completely different explanation for the stable patterns which one can observe in tagging systems exists, namely the regularities and stability of natural language systems. Tagging systems are built on top of natural language and if all natural language systems stabilize over time, also tagging streams will stabilize. Zipf's law [33] states that the frequency of a word in a corpus is proportional to the inverse of its frequency rank and was found in many different natural language corpora (cf. [25]) However, some researcher claim that Zipf's law is inevitable and also a randomly generated letter sequence exhibits Zipf's law [24, 18]. Recent analysis refuted this claim [6, 14] and further showed that language networks (based on word co-occurrences) exhibit small world effects and scale-free degree distributions [15].

8. CONCLUSIONS

Based on an in-depth analysis of existing methods, we have presented a novel method for assessing the semantic stabilization in social streams. We have applied our method to different social tagging streams and to different synthetic tagging streams via simulations. Our results reveal that stability and implicit consensus on the description of resources in social tagging systems cannot solely be explained by the imitation behavior of users; however a *combination of* imitation and background knowledge exhibits highest and fastest semantic stabilization and consensus formation. Summarizing, our work makes contributions on three different levels.

Methodological: Based on systematic investigations we identify potentials and limitations of existing methods for asserting semantic stability in social tagging systems. Using these insights, we present a novel, yet flexible, method which allows to measure and compare the semantic stability in different tagging systems in a robust way. Flexibility is achieved through the provision of two meaningful parameters, robustness is demonstrated by applying it to two random control processes. Our method is general enough to be applicable beyond social tagging systems and we believe it is also useful for analyzing stabilization in other stream based systems such word-streams of the edit history of Wikipedia pages or word-streams of hashtags or URLs.

Empirical: We conduct empirical analysis of semantic stabilization of distinct social tagging streams and natural language streams using our method. We find that the semantic stabilization of tagging streams in systems which support imitation mechanisms goes beyond what can be explained by the semantic stability of natural language and random control processes, while the stability of tagging streams in systems which do not support any imitation show similar semantic stabilization patterns than natural language streams and synthetically generated tag streams drawn from power law tag distributions.

Explanatory: We investigate factors which may explain the stabilization and consensus formation process in social tagging systems using simulations. Our results show that tagging streams which are generated by a *combination of* imitation dynamics and shared background knowledge exhibit faster and higher semantic stability than tagging streams which are generated via imitation dynamics or natural language phenomena alone.

Our findings are relevant for researchers interested in developing more sophisticated methods for assessing semantic stability and agreement in tagging streams and for practitioners interested in assessing the extent of semantic stabilization in social tagging systems on a system scale.

Acknowledgments. We thank Dr. William Webber for assistance with his RBO metric and Dr. Harry Halpin for assistance with his semantic stability measure. This work is in part funded by the FWF Austrian Science Fund Grant I677.

9. REFERENCES

[1] J. Alstott, E. Bullmore, and D. Plenz. Powerlaw: a python package for analysis of heavy-tailed distributions, 2013.

[2] D. Bollen and H. Halpin. The role of tag suggestions in folksonomies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 359–360, New York, NY, USA, 2009. ACM.

[3] C. Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C-Particles and Fields*, 46(2):33–37, 2006.

[4] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, 2007.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[6] A. Cohen, R. N. Mantegna, and S. Havlin. Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, 5(01):95–104, 1997.

[7] K. Dellschaft and S. Staab. An epistemic dynamic model for tagging systems. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, HT '08*, pages 71–80, New York, NY, USA, 2008. ACM.

[8] W.-T. Fu, T. Kannampallil, R. Kang, and J. He. Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(3):12:1–12:37, 2010.

[9] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human*

Language Technologies: short papers - Volume 2, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[10] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[11] O. Görlitz, S. Sizov, and S. Staab. Pints: peer-to-peer infrastructure for tagging systems. In *Proceedings of the 7th international conference on Peer-to-peer systems*, IPTPS '08, pages 19–19, Berkeley, CA, USA, 2008. USENIX Association.

[12] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[13] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA, 2007. ACM.

[14] R. F. i Cancho and B. Elvevåg. Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLoS ONE*, 5(3):e9411+, 2010.

[15] R. F. i Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

[16] C. T. Kello, G. D. A. Brown, R. F. i Cancho, J. G. Holden, K. Linkenkaer-Hansen, T. Rhodes, and G. C. Van Orden. Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232, 2010.

[17] M. E. Kipp and D. G. Campbell. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–18, 2006.

[18] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on*, 38(6):1842–1845, 1992.

[19] N. Lin, D. Li, Y. Ding, B. He, Z. Qin, J. Tang, J. Li, and T. Dong. The dynamic features of delicious, flickr, and youtube. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):139–162, 2012.

[20] J. Lorince and P. M. Todd. Can simple social copying heuristics explain tag popularity in a collaborative tagging system? In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 215–224, New York, NY, USA, 2013. ACM.

[21] G. Macgregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library review*, 55(5):291–300, 2006.

[22] A. Mathes. Folksonomies: Cooperative classification and communication through shared metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`, 2004. Accessed: 2014-01-11.

[23] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15, 2007.

[24] G. A. Miller, N. Chomsky, et al. Finitary models of language users. *Handbook of mathematical psychology*, 2:419–491, 1963.

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

[26] A. Rapoport. *Zipf's law revisited*, pages 1–28. Studies on Zipf's Law. Studienverlag Bockmeyer, 1982.

[27] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Data Science and Classification*, pages 261–270. Springer, 2006.

[28] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, pages 181–190, New York, NY, USA, 2006. ACM.

[29] L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, ESWC '07, pages 624–639, Berlin, Heidelberg, 2007. Springer-Verlag.

[30] C. Wagner, S. Asur, and J. Hailpern. Religious politicians and creative photographers: Automatic user categorization in twitter. In *ASE/IEEE International Conference on Social Computing, SocialCom '13*, 2013.

[31] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, 2010.

[32] G. U. Yule. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. 213(402-410):21–87, 1925.

[33] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[34] A. Zubiaga, C. Körner, and M. Strohmaier. Tags vs shelves: from social tagging to social classification. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 93–102, New York, NY, USA, 2011. ACM.