Emergent Semantics from Game-induced Folksonomies

Lilian Weng* and Filippo Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing, Indiana University, Bloomington

ABSTRACT

We describe the *GiveALink Slider*, an experimental social tagging game designed with the purpose of generating meaningful and useful annotations to improve upon the drawbacks of existing folksonomies. Knowledge, in the form of reliable annotations, is validated and accumulated through the implicit interactions among multiple players. In this paper we explore the hypothesis that such a game can improve both the quality and quantity of social annotation data. Our evaluation of game-induced annotations shows that games may improve on the semantic structure of existing folksonomies from several perspectives, including searchability, novelty, and coherence. Games can therefore play a valuable role in the collection of helpful annotations, by leveraging human power and specific game design features.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Design, Measurement, Experimentation, Human Factors

Keywords

Social tagging, GiveALink, Game with a purpose, Folksonomies

1. INTRODUCTION

Users can collaboratively organize and share Web resources through social tagging, as is widely done in many popular Web 2.0 applications. Various types of online resources can be tagged, reflecting user interests and becoming easily accessible to the public. These resources include Web pages (Delicious), blog posts (LiveJournal), photos (Flickr), music (Last.fm), movies (YouTube), and short messages (Twitter). The output of social tagging systems forms *folksonomies*, or community-induced taxonomies of important and emerging concepts [31, 22]. Folksonomic collections of social annotation data have been shown to be useful to improve social navigation [25], Web search [6, 35], personalized experience [17], recommendation services [21], and social links prediction [28]. However, folksonomies do not scale with the growth rate of the Web [11], so that annotations are sparse. Their reliability is also often questioned, as spammers attempt to manipulate and exploit tagging systems [18].

To mitigate these limitations, this paper explores the hypothesis that *games* can leverage human power to produce high-quality social annotation data, inducing meaningful emergent semantics. We exemplifies with a specially designed game, *GiveALink Slider* (`slider.givealink.org`) [32, 34], to demonstrate the potentiality of crowdsourcing useful folksonomies with games and explore the effectiveness of the mechanism we previously proposed [33]. Our tagging games follow the paradigm of *games with a purpose* [3, 4], solving hard computational problems by asking real people for helpful input in an entertaining way. We further adopt the design principles of the *chain model* [33] for object association games. The idea of the chain model is to collect semantic descriptions of various objects and discover hidden relations among them.

After introducing relevant background in Sec. 2, we discuss game design in Sec. 3, aimed at making the GiveALink Slider entertaining while generating meaningful and useful output. We perform an analysis of data collected over an experimental user study, involving over 34,000 annotations by 200 players, with focuses on patterns of user tagging behaviors (Sec. 4) and evaluations of the quality of game-driven folksonomies (Sec. 5). The report of the game output is the main concentration of this paper, as it presents the value and the quality of game-driven emergent semantics and demonstrates various approaches for social annotation evaluations.

Although our game is still an exploratory project where we resolve to test whether certain design decisions can lead to annotations of satisfactory qualify as a complement to free tagging, the analysis does give substance to the hypothesis: such similar games can play a valuable role in the collection of quality data by leveraging human abilities and specific game design features. The main contributions and findings of the paper are:

- We find that game annotations tend to gravitate toward topics familiar to users (Sec. 4.1).
- The confidence of annotations is validated through the agreement among multiple players, suggesting that game-induced annotations become more reliable over time (Sec. 4.2).
- We show that the patterns of social tagging activity of the game are consistent with human language, and in particular

*Corresponding author. Email: `weng@indiana.edu`

CrowdKDD '12, August 12, 2012, Beijing, China

with both Zipf's law on word frequencies [36] and Heap's law on vocabulary growth [10] in written text (Sec. 4.3).

- Evidence is offered that the game elicits novel annotation, aiding the discovery of semantic relations that used to be hidden (Sec. 5.1).
- Game-driven folksonomies complement the existing social annotation repository, evidenced by a few case studies lending further support for the reliability of the game data (Sec. 5.2).
- Contrary to labeling bookmarks for personal reference, players are found to tag both general and specific resources with general tags, making these resources easier to search (Sec. 5.3).
- Finally, we show that the semantic networks of tags and resources networks become denser, better connected, and more coherent as a result of the game-induced annotations (Sec. 5.4).

2. BACKGROUND

Social tagging data come in the form of *triples*, defined as tripartite relationships $\langle u, r, t\rangle$ between a *user* u, a *resource* r, and a *tag* t [24]. An *annotation* in the paper is defined as a tag-resource pair $\langle t, r\rangle$, aggregated across multiple users who believe that t is a descriptive word for r, and who contribute corresponding triples.

Social tagging data, generated by autonomous behaviors of all the users involved, help people manage and organize their online resource collections, and make sharing and searching easier. Other than these original functions, the data prove to be highly valuable in many other fields. For example, annotations can help enhance Web search [6, 35], generate user profiles [23], personalize recommendations and other user experiences [7, 29], improve social navigation and information accessibility [16, 3, 15], promote semantic Web techniques and construction of ontologies [13]. However, users share annotations largely for individual needs and aspirations, sometimes leading to low-quality annotation data. Current social tagging systems lack enough motivation for the majority of users to label many resources with sufficient numbers of accurately descriptive tags, and the number of new pages posted per day in social tagging systems is small compared to the rate of Web growth [11]. This causes a sparse semantic network of social annotations. Without any control over tagging behaviors, users can easily employ poor tags or even abuse the system by spamming [18]. They can also use tags that are unrelated to a resource, too general to meaningfully describe a given resource, or so specific that they are only useful for one individual.

The idea of using games to help boost both the quality and quantity of social tagging is largely inspired by the popular usage of serious games, social games, and crowdsourcing applications. *Serious games* are games designed with a non-entertainment goal, usually to facilitate learning in education, military, and other settings. *Crowdsourcing* applications harness community knowledge to accomplish certain tasks. Examples include volunteer-based human knowledge repositories such as Wikipedia, human subject task marketplaces such as Amazon Mechanic Turk [14], and scholarly data collection and impact evaluation tools such as Scholarometer [12]. The idea of *games with a purpose* is a combination of serious gaming with crowdsourcing. The best known instance is the ESP Game [3], in which players are paired randomly and asked to achieve agreement on image labels. Many more games with such non-entertainment intentions are utilized in social contexts for both scientific and business purposes. Games have been leveraged to solve protein folding problems in biochemistry [8], celestial object recognition (`galaxyzoo.org`), and human brain

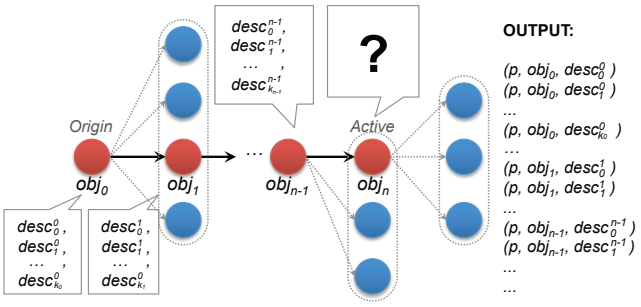


Figure 1: The primary design principle of GiveALink Slider: the chain model for object association games.

training (`www.lumosity.com`). Microsoft uses games in Club Bing (`clubbing.com`) to promote the Bing search engine. All of these applications are instances of *human computation* [4].

The game presented here is designed as a part of the `GiveAlink.org` project, a social bookmarking platform built for research purposes. Previous research in the GiveALink project benefits the game a lot, including a large initial triple repository, a scalable and reliable similarity measure named Maximum Information Path (MIP) [19, 20] for declaring relations among resources, tags, and users, and a social spam filter [18]. The semantic space extracted from the social tagging data is mainly represented as networks of tags or resources, such as the inter-tag correlation graph defined in [9]. Edges can be associated with the similarity values among nodes if the network is weighted. MIP is used for assigning the edge values in the Slider game.

GiveALink Slider originates from the simple idea of building a chain of semantically related objects. The objects are connected by a measure of similarity. The players extend the chain by making these relationships explicit. The idea is formalized as *chain model* [33] for object association games that collect descriptions about, and discover hidden relationships among, Web resources, media, people, and even geographic locations. The chain model, illustrated in Fig. 1, produces an ordered sequence of objects $\langle obj_0, \ldots, obj_n\rangle$. We refer to the last element obj_n as the *active* object. Chain model games allow players to characterize an object obj_i with a set of descriptions $D^i = \{desc^i_0, \ldots, desc^i_{k_i}\}$ in some language. At each step the player p can add a new description to obj_n or make a game *move* to extend the chain, i.e. a transition $obj_n \overset{\mu}{\mapsto} obj_{n+1}$ where μ is a user-defined relation. To facilitate the player's decision of the next object in the chain, the model suggests a set of *candidate* objects C^n that are computed from a system-defined similarity measure with respect to the active object.

To discover user tagging behavior patterns is another related topic concerning the study of user behavior to improve the user experience and support system design. Previous studies have shown that tags are frequently reused even among people who have different background and knowledge [27], and personal experience, such as familiarity with resources, influences tag effectiveness for content sharing [17]. In this paper we will investigate how social tagging works more effectively for sharing and searching, driven by a different tagging goal and by design features particular to gaming. Social tagging can also be exploited to characterize shared user interests [27]. This is analogous to the approach we propose for confirming annotations by shared social tagging actions. In our game, we design a measure of knowledge agreement and track how it changes over time.

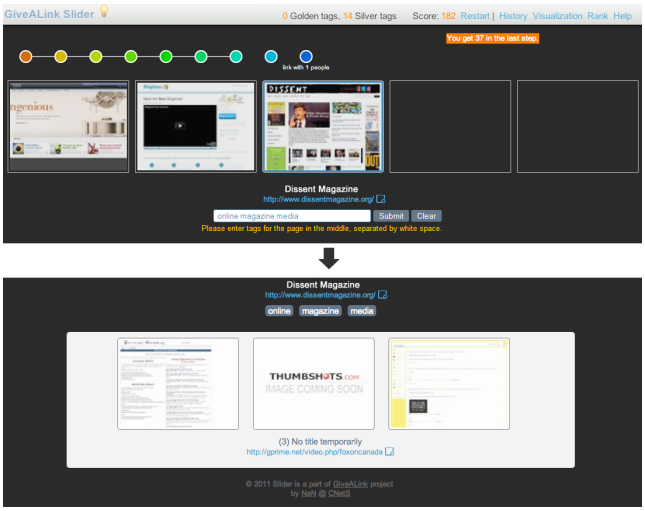


Figure 2: Slider interfaces for tagging the active page and selecting pages among provided candidates.

3. GIVEALINK SLIDER

3.1 Game design

GiveALink Slider (`slider.givealink.org`) is built on the chain model for object association games [33, 34], where players can create chains of relevant Web resources by tagging and selecting Web pages. We experiment several design mechanisms in Slider to foster better game experience and the qualified data collections simultaneously.

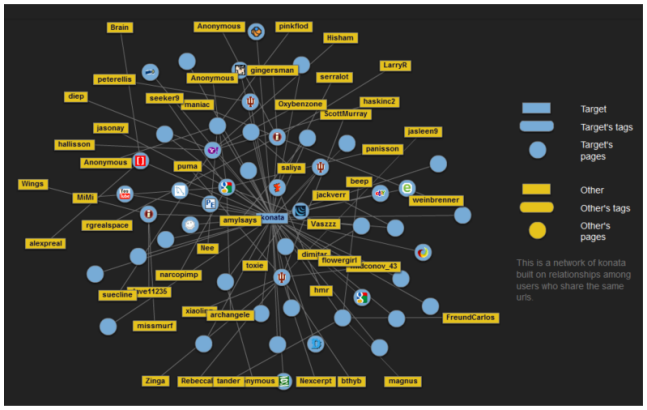
The nodes of the chain model in Slider represent Web resources. Players build chains of pages and generate descriptions by choosing and tagging these pages. The origin page obj_0 in the chain is provided at random, and then the player extends the chain by continuously tagging the active page with one or multiple tags. In the main screen of the game (Fig. 2) the chain is displayed in a carousel, and the player can easily switch among pages. The URL, title, and thumbnail of each page help the player easily learn its content. Each time the player enters tags for a page, the game displays three candidate pages based on the new submission for the player to choose the next node in the chain. Another indicator of the chain is presented as a series of colored nodes above all Web page thumbnails: connected adjacent nodes represent related pages, while nodes may be disconnected due to poor tag inputs or inappropriate candidate choices. It notifies the player whether tags at the previous step is good or not. The final output of the Slider game is a collection of manual labels for various Web pages.

Scoring. We expect to discover novel semantic relations and to strengthen relationships that seem to be weaker than they should be. This goal is fortified with the game scoring system, where each user submission is treated as a set of annotation pairs and each pair is classified into one of three categories:

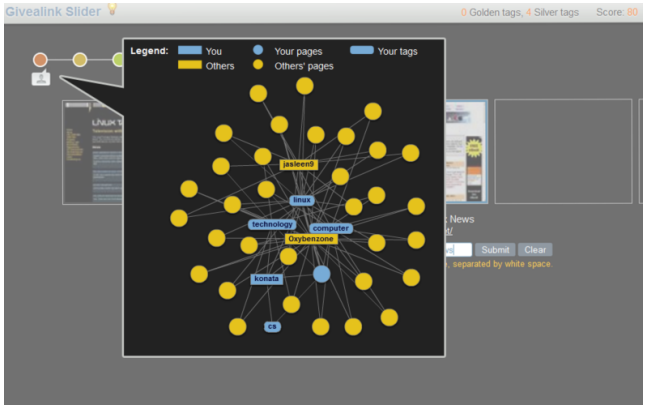
Trusted annotations overlap with data in reliable external sources, such as GiveALink and Delicious;

Novel annotations have been suggested by multiple previous game players, but not found in external data sources;

New annotations appear for the first time, and therefore it is hard to judge whether they are reasonable enough.



(a) Weak Social Links



(b) Strong Social Links

Figure 3: Visualizations of social links in GiveALink Slider.

The reliability of *trusted* and *novel* annotations has been confirmed by external data or other players. We reward players for qualified submissions if the proportion of *trusted* or *novel* annotations in one submission is higher than a threshold ratio. However, the relevance of *new* annotations is still questionable, so they are saved for future verification. The score of each tagging step is obtained by adding the score associated with each annotation pair. *Novel* annotations are most valued and *new* ones are worth less. Unqualified submissions lead to loss of points.

Enjoyability. The basic motivation of the game is to explore unknown Web pages, while players can observe how their minds flow. It is similar to one of the incentives of the online multiplayer word association game, *Human Brain Cloud* [2], that had achieved big success in the past.

To make the game more enjoyable, we visualize social ego networks based on game-driven links among players (Fig. 3). During each tagging step the player may be connected to other players by *strong social links* if they share the same annotations. *Weak social links* can be constructed by sharing either common tags or resources. Both types of social links can be displayed in network visualizations. Another incentive comes from the design of badges of achievement, each corresponding to a game task with multiple levels of difficulty. Completion of harder tasks earns a larger bonus. Tasks deal with Web pages in the chain, social links with other players, score milestones, and other game features.

Flexibility. Additionally, players are allowed to discard pages that are spam, broken links, or written in an unknown language,

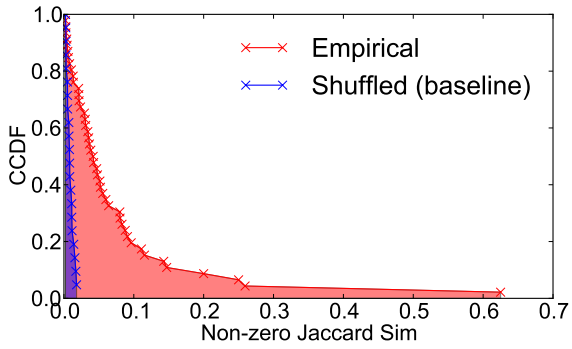


Figure 4: The red area is the complementary cumulative distribution, $P(X > x)$, of Jaccard similarities between tags in GiveALink (base vocabulary) and Slider game, for each player. The blue area is the baseline distribution, generated by shuffled game tags across all game participants.

since those pages distract players and result in poor annotations. The player can replace a candidate page by reporting the reason. All the spam or broken link reports are stored, flagging suspicious or invalid resources and helping improve the accuracy of the classifier employed by the GiveALink spam detector [18].

3.2 User study

We released GiveALink Slider for a six week online user study between March and May 2011. The study eventually involved 200 users. Most participants are of background in computer science, HCI or game design, as we announced the study in several online forums, discussion groups or mailing lists related with these fields. Players generated 43,795 distinct triples (34,642 annotations), among which 13,860 triples (9,111 annotations) were confirmed. On average, each player built a chain of about 26 pages during one game, contributed about 6 chains in total, and each Web page gained about 6 tags.

4. USER TAGGING BEHAVIOR

In this section we analyze the collection of triples generated by the game to study the tagging behavior of players. First, we compare tags from the game versus those from an existing social bookmarking system to examine whether players have a preference for familiar tags. Second, we define a measure of *confidence* to quantify the reliability of an annotation pair. The average confidence as a measure of the confirmed knowledge in the whole system is found to grow with the player involvement and the system size. Third, we compare patterns of tag usage with known regularities of written text.

4.1 Tag familiarity

Learning a new word requires more work than referencing known words. It is reasonable to assume that players may prefer to reuse known words when they need to tag Web resources. If we consider an individual's vocabulary to be a tag repository with size limited by human memory, the tags in the game should to some extent overlap with the tags used previously in other places or for other purposes. Since we have existing bookmark collections from many game players in the GiveALink system, we are able to define all the tags of a player in GiveALink as a *base vocabulary* that she knows and has used before. The Jaccard similarity between the base vocabulary of a player and her game tags represents the extent

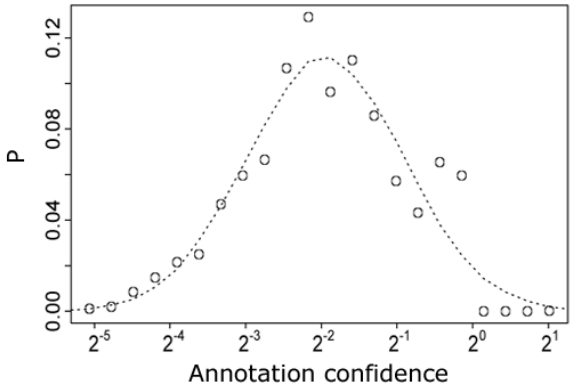


Figure 5: Probability distribution of annotation confidence. The data points are grouped into log-size bins and the frequency of each bin is normalized by the width of the bin. The distribution can be fitted by a log-normal with $\mu = -1.34$ and $\sigma = 0.72$ by maximum-likelihood.

to which the player tends to employ familiar tags. As a baseline, we shuffled game tags across players to simulate players without any memory of words for reference and who thus contribute game tags at random. Fig. 4 shows that similarities are higher than the baseline, suggesting that players are more likely to describe objects according to familiar tags in their vocabulary.

4.2 User agreement

A basic semantic relation in social tagging is an annotation $\langle t, r \rangle$ indicating that a Web resource r can be semantically described by a word t. To determine whether an annotation is reliable (semantically reasonable), which is one of the primary goals of Slider, we can validate it across multiple game participants. Each triple contributor can increase the confidence in the reliability of annotations by agreeing with previous players. Agreement among multiple players on a single annotation is strong evidence that this semantic relation is reliable. To quantify the extent to which players concur on annotation pairs, we define a measure of agreement named *confidence*:

$$c(r,t) = \frac{N(r,t) - 1}{\sqrt{(N(r) - 1)(N(t) - 1)}}$$

where r is a Web resource, t is a tag, and $N(\cdot)$ is the number of unique users with a resource, tag, or annotation. The numerator $N(r,t)$ is the absolute agreement count for annotation $\langle t, r \rangle$. The denominator reduces the influence of very popular resources or tags; a large number of people using an annotation may result from an extremely popular resource or tag, rather than a high confidence in the association between resource and tag. For example, the annotation $\langle$`google.com`, `web`$\rangle$ may have more supporters than the annotation $\langle$`google.com`, `pagerank`$\rangle$, owing to the fact that the tag `web` is heavily utilized to tag any online resource. Consequently, users may tag `google.com` with `web` simply because the tag itself is popular and general rather than because the tag is thought to be especially relevant. The presence of both tag and resource popularity in the denominator of the definition of confidence discounts such a popularity bias. All the counts are decremented by one to avoid personal use cases that has not be validated by more than two distinct players.

The confidence of a pair $\langle t, r \rangle$ represents the strength of the con-

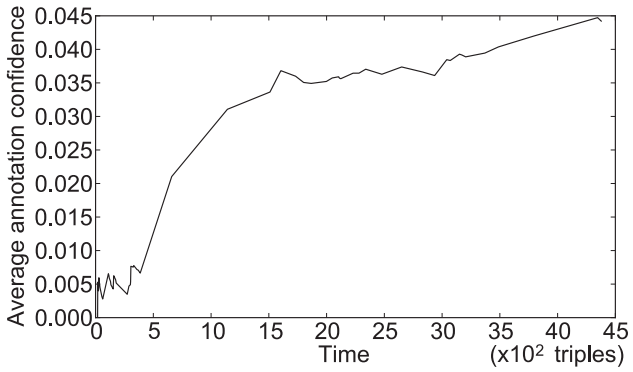


Figure 6: Growth of average confidence with system time, measured by the total number of triples.

nection between tag t and resource r, that is, how well t describes r. We consider an annotation with high confidence to be reliable. According to Fig. 5, most annotation confidence scores are between $2^{-3} = 0.125$ and $2^{-1} = 0.5$. We observe only a few annotations with extremely low or high confidence scores. The low-confidence annotations could be too personal to be useful to a general audience and the high-confidence ones are likely well known, so the most valued knowledge derives from the social annotations with intermediate confidence.

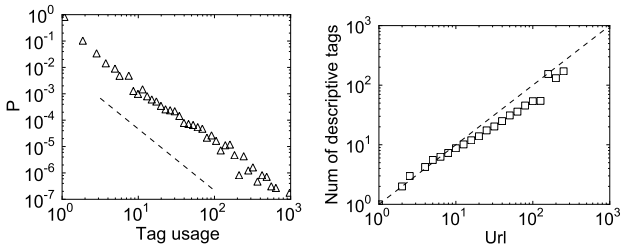
With more players contributing triples to the system, the number of confirmed semantic relations increases and the annotations gain confidence. As shown in Fig. 6, the average annotation confidence grows with system time, measured by the total number of available triples. This also suggests that confidence is a good measure of annotation reliability, and could even be used to gauge the total amount of semantic knowledge accumulated by the system. The Slider game is able to capture more valued knowledge with more accumulated triples. In Web applications with social tagging components, confidence can be incorporated into search algorithms as it scores the capability of a keyword to describe search results.

4.3 Tagging versus written text

The way people collectively describe Web resources in the Slider game is consistent with regularities of word usage in written text [5]. We confirm this by verifying that both Zipf's law [36] and Heaps' law [10] hold for game-induced social tagging data. Zipf's law says the global word frequency in written documents is inversely proportional to word rank by frequency, which is equivalent to saying that the frequency has a power-law distribution with exponent close to 2. Heaps' law describes the sub-linear growth of the number of distinct words with the length of a document, according to which vocabulary size grows more slowly than document size.

We can verify both regularities for the game output if we consider each Web resource as a document associated with a set of tags from the annotations. Under this view, the global popularity of tags displays a power-law distribution consistent with Zipf's law (Fig. 7(a)). Furthermore, as the number of tags associated with a single resource increases, the number of distinct tags grows sub-linearly, consistently with Heaps' Law (Fig. 7(b)). Agreement on the same annotation by multiple players generates the repetitions of tags. Therefore Heaps' Law is a reflection of the rise of system confidence and user agreement.

5. DATA QUALITY



(a) Zipf's Law (b) Heaps' Law

Figure 7: (a) Distribution of tag frequencies. The triangles are empirical data from the game, and the dashed line marks the slope corresponding to a power law with exponent around 2. (b) Number of unique tags versus total number of tags used for a single URL. The squares represent empirical data from the game, and the dashed line is the diagonal, showing sub-linear growth.

We evaluate the quality of game-driven folksonomies from several perspectives, including the novelty of newly collected data, the reliability of semantic descriptions, resource searchability, and semantic network coherence. Game-driven social tagging data improves on each of these aspects, suggesting that it is a good complement to the semantic space of existing social tagging data.

5.1 Novelty

One goal of game-driven folksonomies is to enhance the social tagging data with fresh semantics that are missing in existent systems. We are able to accumulate new triples and discover hidden semantics only when the game data does not overlap excessively with existing data. First, we examined the novelty of annotations by calculating the overlap between the game output and existing social tagging data. There are around 9,000 annotations confirmed through Slider; only 5.74% of them overlap with the GiveALink bookmark repository and 21.87% appeared through the Delicious tag recommendation API. The remaining majority are novel annotations discovered via the game.

Then we analyzed the implicit semantics hidden under the collected triples and annotations, as semantic connections among tags or resources. We calculated the similarity between each pair of tags or Web resources based on collected triples. We leveraged the maximum information path (MIP) similarity [20], a scalable and collaborative measure, to compute relationships among all tags or resources, and generated two sets of tag-tag and resource-resource similarities, respectively. The MIP similarity measure has the following definition:

$$\sigma_u^{MIP}(x_1, x_2) = \frac{2\log(\min_{y \in X_1^u \cap X_2^u}[p(y|u)])}{\log(\min_{y \in X_1^u}[p(y|u)]) + \log(\min_{y \in X_2^u}[p(y|u)])}$$

where x represents a tag or a resource and X is its vector representation [20]. If x is a tag, y represents a resource, and for any user u, $y \in X^u$ indicates that there is a triple $\langle u, y, x\rangle$ where u tags y with the tag x. The similarity has a value between $[0, 1]$ where a similarity of zero indicates that two tags or resources are unrelated. A relation between a pair of tags or resources is considered to be *novel* if $\sigma = 0$ according to existing data but $\sigma > 0$ according to game data. In other words, the relationship is newly discovered through the game. In the data generated by our user study of the Slider game, 60.22% of tag-tag relations and 72.92% of resource-resource relations are novel.

The fact that the game can efficiently accumulate novel triples

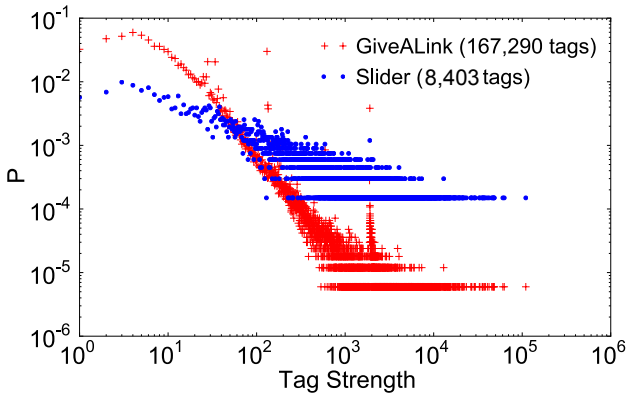


Figure 9: Distribution of tag generality in Slider and GiveALink.

and uncover novel semantics originates from the different goals of the game compared to traditional social bookmarking systems. Most existing social tagging websites mainly function as online bookmark repositories where users are allowed to freely tag resources or upload bookmarks from their browsers. Such annotations primarily reflect information management of an individual user. The game, on the other hand, asks players to tag resources not for the sake of an individual's future reference, but simply for fun, while their mind flows. As a consequence, we are able to avoid overly personal tags and collect more representative and descriptive relations [34].

5.2 Reliability

We wish to find support for our hypothesis that the game mechanisms lead to collection of reliable data. Several tag similarity ego-networks are shown in Fig. 8 to illustrate the good quality of game tags. The comparison between the game and GiveALink suggests game-induced folksonomies have positive effect on the existing social annotation repository, that can be similarly extended to other popular social tagging systems, such as Delicious and Flickr.

Fig. 8(a) displays the top 30 tags of both game and GiveALink. They share several top buzzwords including common internet terms (`internet`, `web`, `software`), suspicious/spam words (`free`), and news related tags (`news`, `business`). Among other non-overlapping tags, Slider shows a preference for technology and education related terms, while the GiveALink tags tend to be broader or more suspicious/spam in nature. Fig. 8(b) displays a tag similarity networks formed by tags similar to the word *social*. Nodes with edges of both colors reflect many reasonable social-related concepts. However, several nodes connected by GiveALink edges (blue), are not clearly relevant to *social*, including a clique of name-like tags (`amanda`, `derek`) and unrelated concepts (`dior`). Top tags in Slider incorporate many popular social networking websites (`facebook`, `linkedin`, `twitter`) and social activities (`sharing`, `tagging`, `gaming`). In general the game tags are of better quality and more searchable. Similar trends can be found in Fig.s 8(c) for *apple* and 8(d) for *sports*.

5.3 Searchability

Halpin *et al.* [9] discuss the *information value* of a tag, which refers to the information conveyed by the natural language term used in the tag and how this makes the tag useful for the retrieval of resources. More precisely, the information value of a tag is interpreted as a function of the number of resources a tag can retrieve

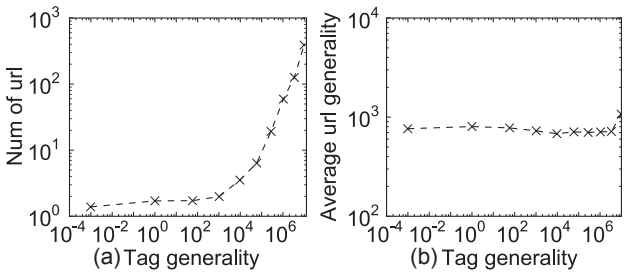


Figure 10: Attractiveness of tags with different generalities.

while searching. Tags with high information value attract future usage from users while contributing to discriminating among resources to make them more searchable and accessible.

The retrieval power of a tag is closely associated with its generality, or capability to describe different resources. When a tag relates to a broad semantic spectrum, it may be adopted for resources in many categories. GiveALink tags are mainly collected from users' browser bookmarks, so some of them are only meaningful for one individual. Those tags can only describe personal resources from personal perspectives, causing disconnected resources and sparse semantic networks; they cannot contribute much to retrieval.

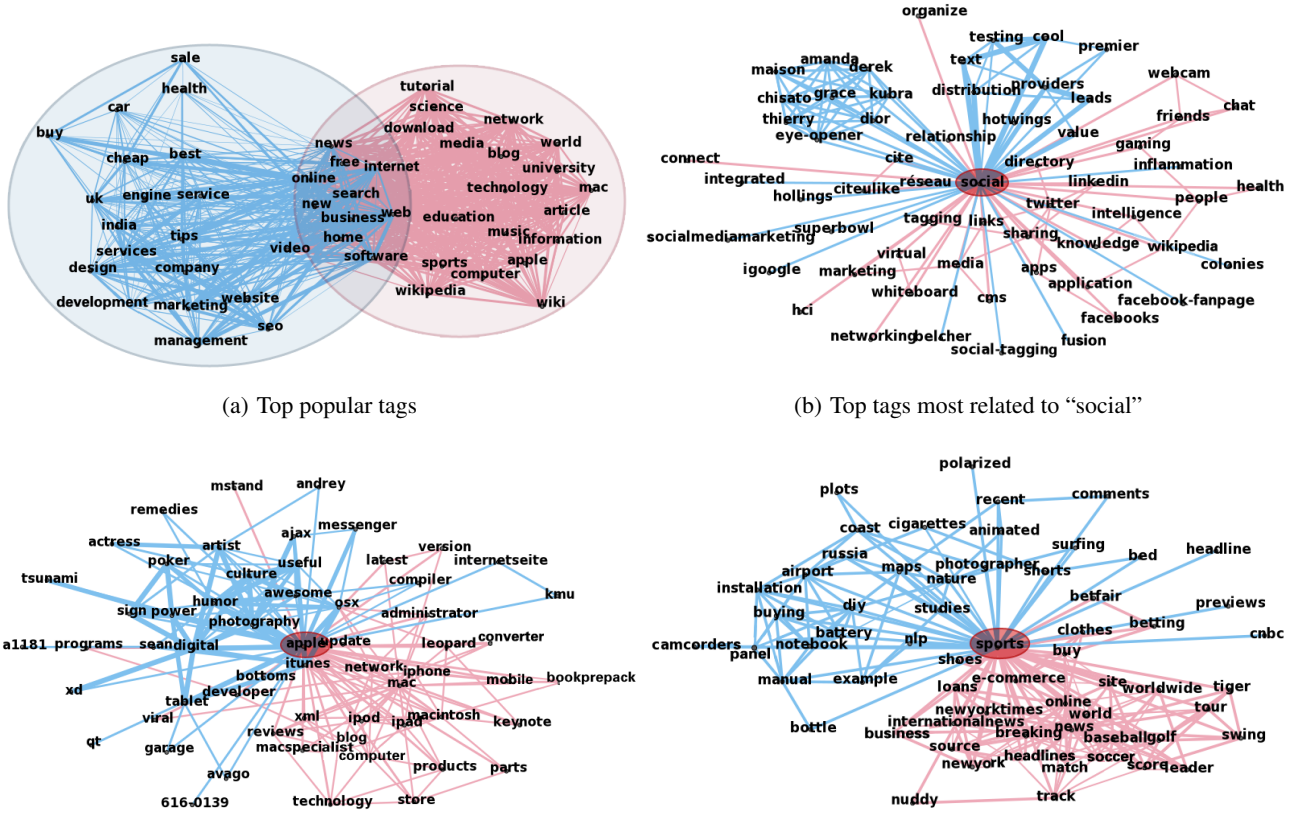
To compare the information value carried by GiveALink and Slider tags, let us quantify the information value of a tag as the sum of its similarities with all other tags, known as *tag generality* or *tag strength*:

$$s(t) = \sum_{t' \in T, t \neq t'} \sigma(t, t')$$

where t, t' are tags, T is the set of all tags, and σ is the similarity measure between a pair of tags. The strength is computed globally for all existing tags in GiveALink, of which Slider tags are a subset. Then we are able to compare the strengths of these two sets of tags. According to Fig. 9, the Slider game succeeds to avoid many tags with very small strength that are too specific to be useful for guiding other users or helping the search. The game therefore leads to more re-usable tags, and to a more "information-rich" semantic space. Note that Fig. 9 is plotted in the logarithm scale; Thus, the majority of the game tags are still specific and novel (Sec. 5.1).

Fig. 10(a) shows a clear trend that general tags are utilized more frequently for describing resources in the game. This is not surprising, considering that general tags are better capable of labeling concepts. On the other hand, similarly to the definition of tag strength, each resource can be associated with a strength or generality value by summing up its similarity scores to all other resources. High-strength resources are widely connected with others by shared annotations, and they usually have general information that is appropriate for a broad group of users. Conversely, low-strength resources are valuable for smaller groups. Thus it is possible to hypothesize that high-strength resources attract high-strength tags, since they both cover broad categories of information and thus they are more likely to be associated with each other. However, as shown in Fig. 10(b), both general (high-strength) and specific (low-strength) resources are annotated with a mixture of high-strength and low-strength tags. This implies that both general and specific resources have equal chances to be tagged with searchable descriptions. While general tags are more popular, they are used to describe specific resources as well, helping users find them in broad categories.

5.4 Discovering hidden semantics



(a) Top popular tags

(b) Top tags most related to "social"

(c) Top tags most related to "apple"

(d) Top tags most related to "sports"

Figure 8: Visualization of tag similarity ego-networks. Red edges are similarity connections from the game, blue ones are from GiveALink.

Since the game can collect novel descriptions (Sec. 5.1), the semantic space of folksonomies is expected to reveal more novel relations and to strengthen weak semantic relations that should be stronger. We can represent the semantics as either unweighted or weighted similarity network, and then study how the space is reshaped with both representations while adding game data. First, the space is found to become denser and better connected when treated as an unweighted network. Second, with consideration of similarity values and edge weights, we illustrate that the weighted network built on folksonomies grows more coherent via the game.

5.4.1 *Network structure*

With more game data flowing into the semantic space, we expect the structure of the folksonomy built from the tripartite relationships between tags, users, and resources to exhibit some changes with more hidden semantics recovered. Semantic spaces of both tags and resources can be created, with each simply being treated as an undirected and unweighted similarity network, where each node represents a single tag or resource and a connection (edge) between a pair of nodes is implied if the similarity is larger than zero. A pair of nodes without a connected edge indicates that either they are not semantically related, or the relation is still unknown to the system.

We built two similarity networks with the same set of 3,000 tags sampled randomly from the intersection of GiveALink and Slider. One network included only relations in the original GiveALink folksonomy, the other also included relations from the game. A comparison between the two networks reveals how the game-driven folksonomies change the semantic space. The same comparison was performed with resource similarity networks. Table 1 summarizes the results of several measures, which clearly demonstrate that with game data included, both tag and resource similarity networks become denser and better connected, with more closed triangles and fewer small, isolated components.

5.4.2 *Metric coherence*

Tag and resource networks can be created with weighted edges if the similarity values are taken into account. We inspected the changes in *coherence* among edge weights based on the hypothesis that the network can evolve to be more coherent because weak relations that should be stronger are recovered by game data.

When we try to describe relations among a set of nodes V with some measurement, such as a distance function $d : V \times V \to \mathbb{R}$, we are able to claim the function d is *metric* if d satisfies non-negativity, identity of indiscernibles, symmetry, and triangular inequality. The first three are trivial. The triangular inequality states that for $\forall a, b, c \in V$, $d(a, b) \leq d(a, c) + d(c, b)$, indicating that the shortest path between a pair of nodes should be the direct link. When given a similarity measure instead of a distance, the analog of the triangular inequality is that the direct similarity between two nodes should be higher than the indirect similarity through a third node. If the direct similarity between a pair of nodes is lower than some indirect similarity, we say that the triangular inequality is violated by the pair and that the space is *semi-metric* [26]. For

Table 1: Semantic network comparisons

	Tag Similarity Network			Resource Similarity Network		
	GiveALink	Combined	Change	GiveALink	Combined	Change
Nodes	3,000	3,000		3,000	3,000	
Edges	147,182	159,247	+8.2%	213,498	307,363	+44.0%
Density	0.0044	0.0048	+9.1%	0.0083	0.0119	+43.4%
Closed triangles	5,090,467	5,657,185	+11.1%	13,684,666	18,383,568	+34.4%
Components	5,917	5,316	–10.2%	4,424	4,199	–5.1%

Table 2: Network metric coherence ratio

Network	GiveALink	Slider	Combined	Change
Tag	0.0963	0.8375	0.1049	+8.87%
URL	0.8908	0.8956	0.9574	+7.48%

instance, if a is similar to b and a third node c is similar to both, but $\sigma(a,b) < \sigma(a,c)\sigma(c,b)$, we say that the pair (a,b) is semi-metric. Finding semi-metric pairs is valuable for discovering novel semantics because they highlight cases where a semantic relationship is implied by transitive closure of the global folksonomy, but has not yet been uncovered by explicit annotations [30].

A network with many semi-metric pairs is not as semantically coherent as a metric network. Our hypothesis is that game data can make the network structure more coherent by increasing metric pairs, since it can strengthen the weak relations that should be stronger according to the triangular inequality. We evaluate coherence by measuring the ratio of the number of metric pairs to the number of all pairs. Using the similarity function σ, let us reformulate the triangular inequality as follows: $\forall a,b,c \in V, \sigma(a,b) \geq \sigma(a,c)\sigma(c,b)$, then the edge (a,b) of the triangle (a,b,c) is counted as being metric.

As reported in Table 2, the metric coherence of both tag and resource similarity networks is improved by the game; game-driven social tagging data makes the GiveALink folksonomy more coherent semantically. It is noteworthy that the resource similarity network presents a much higher metric coherence. This is because the meaning of a tag is more likely to be ambiguous compared to the meaning of a Web resource, thus it is easier to describe the similarity relations among resources in a coherent way.

6. CONCLUSIONS

User-generated social annotations are valuable for many Web applications, such as those providing a better Web search experience or personalized recommendations. However, in current social tagging systems, users can easily employ poor tags or even abuse the system by tagging spam. The lack of control yields tags that are overly personal or general as well as spam. We propose games with the purpose of improving both the quantity and quality of social annotation data. We presented one such experimental social tagging game, the GiveALink Slider, that we have designed and developed for desktop browsers. The basic design principles are derived from the chain model for games that generate object descriptions by object associations [33]. Players win points by extending a chain of related Web pages through tag descriptions. The scoring mechanism fosters novel but accurate tags and the social connection component increases the game's enjoyability.

The output of the game accumulates as more participants get involved. While playing, users label personal bookmarks and other resources with tags. Compared with data from traditional social tagging systems, users tend to apply more general and information-rich words, which is encouraged by easily winning points and building social links among players according to the game's features. Such tags are equally applied to general and specific resources, increasing the searchability and accessibility of both types. The tag space of the game's output is therefore more re-usable. Meanwhile, players prefer familiar words, such as tags previously used for bookmark annotations, since they have limited vocabularies and learning new words is usually more difficult than referencing an old one. *Confidence* is defined for an annotation pair $\langle t, r \rangle$ to represent its reliability from the community's perspective. The value of confidence is proportional to the number of players who describes resource r with tag t. In other applications with social tagging components, confidence can be incorporated into search. The growth of average confidence in game annotations over time validates the definition of confidence as a measure of confirmed knowledge.

According to our evaluation, the quality of game-induced folksonomies meets our expectation of improving exiting social tagging dataset. A large proportion of the game's annotations reveal novel relations among tags or resources, unknown to GiveALink or Delicious. The output difference between our game and other social tagging systems originates from the primary goals of the two applications: one is to label sites for future personal references, the other is simply for fun. Additionally, resources become more searchable because of different user tagging behavior concerning tag choices. While the semantic quality of the data can be sustained by the game mechanisms, the semantic networks of concepts that emerge from these annotations grow denser and better connected as novel relations are absorbed to recover previously hidden semantics. Finally, the semantic networks become more coherent as game annotations are added.

In summary, the result of our study is a small step in entertainingly crowdsourcing large streams of high quality annotations, but it is quite promising. Our game substantiates the hypothesis that user tagging behavior induced by the game rules enriches the quality and quantity of annotations. To solve the scalability problem, a simplified version of game rules is necessary. Another challenge that is yet to be pursued is to design mechanisms that make the games sufficiently entertaining to engage players for long periods of time. More designs to enhance the pleasurableness of gaming need to be adopted, for example, an interesting visualization of word associations by using `d3.js` chord diagram [1]. Besides the Slider, we are designing and developing another tagging game for mobile devices, called *Great Minds Think Alike* (`greatminds.givealink.org`). The lessons learned from the present study will inform the design of this and other games, and it will be challenging to see how they can be applied to deal with other types of objects, such as rich media, locations, and friends.

Acknowledgments

We are grateful to Rossano Schifanella for assistance in the game design and implementation and the members of the Networks and agents Network (`cnets.indiana.edu/groups/nan`) at the

Indiana University School of Informatics and Computing for helpful suggestions during game design and testing. This work is funded by NSF award IIS-0811994 *Social Integration of Semantic Annotation Networks for Web Applications*.

7. REFERENCES

[1] Diagram in `d3.js`, `mbostock.github.com/d3/ex/chord.html`.

[2] Human brain cloud: Massively multiplayer word association game, `www.readwriteweb.com/archives/human_brain_cloud.php`.

[3] L. V. Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, 2004.

[4] L. V. Ahn and L. Dabbish. Designing games with a purpose. *Communication of the ACM*, 51(8):58–67, 2008.

[5] M. Ángeles Serrano, A. Flamminia, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE 4(4):e5372*, 2009.

[6] S. Bao, X. Wu, B. Fei, G. Xue, Z. Su, and Y. Yu. Optimizing web search using social annotation. In *Proc. of Intl. World Wide Web Conf. (WWW)*, 2007.

[7] F. Carmagnola, F. Cena, O. Cortassa, C. Gena, and I. Torre. Towards a tag-based user model: how can user model benefit from tags? In *Proc. of Conf. on User Modeling*, pages 445–449, 2007.

[8] S. Cooper, A. Treuille, J. Barbero, A. Leaver-Fay, K. Tuite, F. Khatib, A. C. Snyder, M. Beenen, D. Salesin, D. Baker, and Z. Popović. The challenge of designing scientific discovery games. In *Proc. of Intl. Conf. on Foundations of Digital Games (FDG)*, 2010.

[9] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proc. of Intl. World Wide Web Conf. (WWW)*, 2007.

[10] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Orlando: Academic Press, 1978.

[11] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. of ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, 2008.

[12] D. T. Hoang, J. Kaur, and F. Menczer. Crowdsourcing scholarly data. In *Proc. of Web Science Conf.*, 2010.

[13] H. L. Kim, A. Passant, J. Breslin, S. Scerri, and S. Decker. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Proc. of IEEE Intl. Conf. on Semantic Computing*, 2008.

[14] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, 2008.

[15] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[16] K. G. Lawson. Mining social tagging data for enhanced subject access for readers and researchers. *Journal of Academic Librarianship*, 35:574–582, 2009.

[17] C. S. Lee, D. H.-L. Goh, K. Razikin, and A. Y. Chu. Tagging, sharing and the influence of personal experience. *Journal of Digital Information*, 10(1), 2009.

[18] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *Proc. of Intl. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2009.

[19] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proc. of Intl. World Wide Web Conf. (WWW)*, 2009.

[20] B. Markines and F. Menczer. A scalable, collaborative similarity measure for social annotation systems. In *Proc. of ACM Conf. on Hypertext and Hypermedia*, 2009.

[21] B. Markines, L. Stoilova, and F. Menczer. Social bookmarks for collaborative search and recommendation. In *Proc. of National Conf. on Artificial Intelligence (AAAI)*, pages 1375–1380. AAAI Press, 2006.

[22] C. Marlow, M. Naaman, danah boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. of ACM Conf. on Hypertext and Hypermedia*, 2006.

[23] E. Michlmayr and S. Cayzer. Learning user profiles from tagging data and leveraging them for personal (ized) information access. In *Proc. of Intl. World Wide Web Conf. (WWW)*, 2007.

[24] P. Mika. Ontologies are us: A unified model of social networks and semantics. *The Semantic Web – ISWC 2005*, 5:522–536, 2005.

[25] D. Millen and J. Feinberg. Using social tagging to improve social navigation. In *Workshop on Social Navigation and Community-based Adaptation*, 2006.

[26] L. M. Rocha. Semi-metric behavior in document networks and its application to recommendation systems. In *Soft Computing Agents: A New Perspective for Dynamic Information Systems*, pages 137–163. IOS, 2002.

[27] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu. Individual and social behavior in tagging systems. In *Proc. of ACM Conf. on Hypertext and Hypermedia (HT)*, 2009.

[28] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: Social link prediction from shared metadata. In *Proc. of ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, 2010.

[29] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proc. of Conf. on Recommender Systems*, 2008.

[30] L. Stoilova, T. Holloway, B. Markines, A. Maguitman, and F. Menczer. Givealink: Mining a semantic network of bookmarks for web search and recommendation. In *Proc. of SIGKDD Workshop on Link Discovery: Issues, Approaches and Applications*, 2006.

[31] T. V. Wal. Folksonomy coinage and definition. http://vanderwal.net/folksonomy.html.

[32] L. Weng and F. Menczer. Givealink tagging game: An incentive for social annotation. In *Proceeding of the SIGKDD Workshop on Human Computation Workshop*, 2010.

[33] L. Weng, R. Schifanella, and F. Menczer. The chain model for social tagging game design. In *Proc. of Intl. Conf. on the Foundations of Digital Games (FDG)*, 2011.

[34] L. Weng, R. Schifanella, and F. Menczer. Design of social games for collecting reliable semantic annotations. In *Proc. of IEEE Intl. Conf. on Computer Games (CGAMES)*, 2011.

[35] S. Xu, S. Bao, Y. Cao, and Y. Yu. Using social annotations to improve language model for information retrieval. In *Proc. of ACM Conf. on Information and Konwledge Management*, pages 1003–1006, 2007.

[36] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Cambridge: Addison-Wesley, 1949.