Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2013 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-18-2013

Don't Look Here: Off-Limits Words Bias Play in the ESP Game

David Bodoff
University of Haifa, dbodoff@uiv.haifa.ac.il

Eran Vaknin
University of Haifa, eran@jordanvalley.com

Recommended Citation

Bodoff, David and Vaknin, Eran, "Don't Look Here: Off-Limits Words Bias Play in the ESP Game" (2013). *PACIS 2013 Proceedings*. 183.
http://aisel.aisnet.org/pacis2013/183

DON'T LOOK HERE: OFF-LIMITS WORDS BIAS PLAY IN THE ESP GAME

David Bodoff[1], University of Haifa, Graduate School of Management, Haifa, Israel, dbodoff@uiv.haifa.ac.il

Eran Vaknin, University of Haifa, Graduate School of Management, Haifa, Israel, eran@jordanvalley.com

Abstract

Social computation is a paradigm in which a software application supports social interaction, but whose real purpose is the data trail that is left by that interaction, such as tags, recommendations, and so on. We explore a possible imperfection in the data that is generated by these games. Specifically, we investigate whether the data generated by previous participants influences the data that is generated by subsequent participants. We investigate this in the context of the ESP game. A feature of the ESP game is that words that have been generated by previous players are off-limits to subsequent players. The idea of this feature is to ensure that the system accumulates a variety of different words for each image. We consider the possibility that ironically, players are actually biased to suggest words that are related to the taboo words themselves. Based on anchoring and priming theories, we predict that the words players suggest will be related to the taboo words, and that this phenomenon will limit the variety of words that are collected for a given image. An empirical experiment confirms these predictions. This effect threatens to limit the potential value of socially generated information in many applications including recommender systems and ESP-like tagging systems, where later contributors are exposed to the inputs provided by earlier contributors.

Keywords: ESP game, social computing, human computation, coordination games.

[1] whose research was supported by the Seoul National University Foreign Faculty Foundation

1 INTRODUCTION

Many kinds of computer systems facilitate communication in addition to computation. But the connection between social interaction and computer systems is further crystallized in the emerging paradigm of ‘social computing’. Social computing involves an interplay between people’s social behaviors and their interactions with computers (Dryer, Eisbach et al. 1999). It includes applications such as social networks and wikis (Muthén 1998-2011). A related concept is ‘human computation’, whose main idea is to use humans to perform computations. The two paradigms -- social computing and human computation -- coincide in applications that facilitate social interaction, which produce valuable information as a by-product. These by-products take the form of metadata, opinions, recommendations, evaluations, interpretations, etc. that may be useful to others. Some well-known commercial applications of this kind are Flickr, Wikipedia, Digg, Facebook, etc. We coin the term “social computation” to refer to applications that combine elements of both human computation and social computing. In some cases, the data that is generated is numerical, e.g. average rating. But much of the data is also textual, e.g. tags and comments, for the simple reason that people communicate with one another mostly via words.

In this paper, we explore a bias in the data that is generated through social computation. In particular, we show how the data that is generated by previous users influences the data that is generated by subsequent users. This results in a rich-get-richer effect, which limits the variety of information that is ultimately collected. We study this phenomenon in the context of an application called the ESP game, but it is equally applicable to all applications such as collaborative filtering (Konstas, Stathopoulos et al. 2009), folksonomies (Golder and Huberman 2006; Schmitz et al. 2007), FlickR (Sigurbjörnsson and Zwol 2008), and other applications, where users see the words that were already contributed by previous players.

As social computation grows as a paradigm, it will become increasingly important to understand the characteristics – including the limitations and shortcomings -- of the aggregate data that emerges from these applications. Our study aims to establish a theoretical basis for understanding the bias that may enter data that is generated via social computing, as a basis for recommending interventions to correct it.

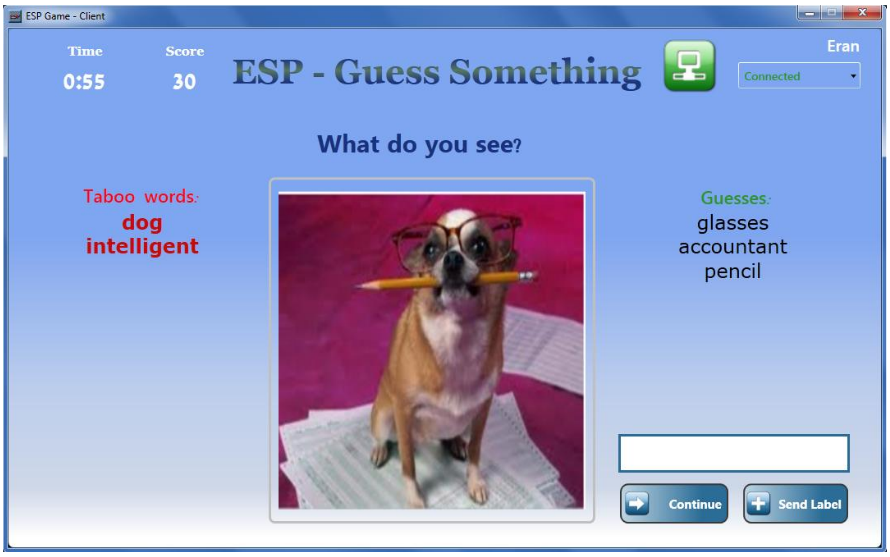
2 BACKGROUND AND MOTIVATION

2.1 Social Computation

Human computation is a growing paradigm (Quinn and Bederson 2011). It “treat[s] human brains as processors in a distributed system, each … perform[ing] a small part of a massive computation” (von Ahn 2006 p. 96), emphasizing that humans are used in place of a computer. A related paradigm is social computing, which emphasizes the connection between social interactions and computing (Quinn and Bederson 2011). Many applications such as Flickr, Wikipedia, Digg, Facebook include both aspects, in that data is produced as a side-effect of social interaction. We introduce the term “social computation” to refer to such applications.

On example of social computation, which we will consider in detail in our experiment, is the ESP game (von Ahn and Dabbish 2004). Assigning labels to images is difficult for computers, and tedious for humans. The ESP game makes it fun, so that human players will do it voluntarily. In the game, each player is assigned to an anonymous partner. The two players cannot communicate with one another. The system shows them both the same image. Each player suggests as many words as he/she likes as labels for that image, until the player and his/her partner have both suggested the same word. As explained by von Ahn (2006), “Once both players have typed the exact same string, a new image appears; they don’t have to type the string at the same time, but each must type the same string at some point while the image is onscreen” (von Ahn 2006 p. 96). The game is normally played under a time pressure, with players trying to match on as many pictures as possible within the allotted time.

Figure 1 shows an example screenshot. On the right side, we can see that this player has so far suggested the words “pencil”, “accountant”, and “glasses” for the picture shown (the left side “Taboo words” are discussed below). At the same time, the partner will have been suggesting words. Evidently, his/her partner has so far been suggesting different words, so they are still working on this picture; once they overlap in any word, that will be recorded as a match and the two players will proceed to the next picture.



Time
0:55

Score
30

Taboo words:
dog
intelligent

Guesses:
glasses
accountant
pencil

Figure 1. Example Image in ESP Game

Although *players*’ goal is to match their partner and win points, the goal of the website deploying the game is its side-effect of image labels. For example, a website such as Flickr or Google may deploy an ESP-like game – as Google did – to collect labels on images. When a pair of players in the ESP game matches on a word, they conclude that that word is evidently a good label for that image, and index the image with that word. Then, when a user of a search tool (which has no direct connection to the ESP game) submits a keyword search for images, the system can look for images with that label. Google deployed the ESP game from 2006 – 2011 under the name ‘Google Image Labeler’. The same idea has been applied to other related applications such as music annotation (Law, Ahn et al. 2007). We choose the ESP as an example of social computation, but the phenomenon we study applies to many if not all social computing applications. In particular, we study how previously generated data influences subsequently generated data.

2.2 Data Quality: Individual Correctness and Aggregate Variety

The purpose of social computation applications – often these applications are in the form of games (Von Ahn and Dabbish 2008) -- is to produce data. When the games are designed, explicit attention is therefore given to introducing features that will ensure data quality. There are two aspects of data quality, i.e. the correctness of each individual piece of data, and the variety of the resulting data set. For an individual piece of user-generated data, a most basic aspect of quality is its correctness (when there is an objective truth) or appropriateness (where there isn’t any objective truth). For example, in CAPTCHA games (Von Ahn, Maurer et al. 2008) in which the user is asked to correctly read an unclear word, there is a single correct answer, while in the Verbosity game (Von Ahn, Kedia et al. 2006) for eliciting commonsense knowledge, or the ESP game for eliciting image labels (von Ahn and Dabbish 2004), there isn’t. Different games ensure this kind of quality with different design features, which are characterized as symmetric or asymmetric (Chan, King et al. 2009). At the aggregate level of analysis, there is a concern with the variety of data that is collected (Von Ahn and Dabbish 2008). This concern is relevant in settings in which there are many possible “right answers”, and one would ideally like to collect as many of these as possible. This is often measured as the number of different tags that are collected (Huang et al. 2010; Ho and Chen 2009; Ho et al. 2010), sometimes called tagging exhaustiveness (Sigurbjörnsson and Zwol 2008).

"Taboo" words is a design feature that has been used in a number of games to help ensure such variety. In the context of the ESP game, the taboo words feature means that after a number of player pairs achieve a match by suggesting a given label for a particular picture, that word becomes off-limits – "taboo" – to future player pairs who encounter that same picture. In this manner, subsequent player pairs are forced to try other labels. This device is intended to ensure that as its by-product, the game generates a variety of labels for each picture. Returning to Figure 1, there were two taboo words in effect – 'dog' and 'intelligent' -- when the depicted user was playing; the user (and his/her partner) were prohibited from entering these words. There would be no particular penalty from trying to enter them, but they would not be considered, and the users would have wasted their limited time.

2.3 Research Questions

In light of the above, an over-arching question facing the social computation paradigm is whether the data it produces can achieve satisfactory levels of quality; empirical results have been mixed (Goh, Ang et al. 2011). Perhaps a more profitable question is not "how good" the data is, but how it can be improved. This requires a specific understanding of the imperfections that the data contains, and the processes that lead to these imperfections. This is the approach taken here.

Ho et al. (2009) and Weber and Robertson (2008) both raise the possibility that in the ESP game, the taboo words – which represent the successful word matches of previous players -- may bias future players' word suggestions. As an example, the user in Figure 1 may be influenced by the taboo words "dog" and "intelligent". At a general level, this possibility means that the data generated by previous players may influence the word suggestions of later players. At a more specific level, it means that – paradoxically – the game feature that was designed to ensure variety in the data, actually introduces a bias that also limits the data variety. This leads to our research questions:

RQ1: In social computation applications that show users the data generated by previous players, are the users influenced by the prior data?

RQ2: If so, does this influence limit the variety of words that is generated when the game is played across a whole population of players?

We consider the concrete example of the ESP game taboo words feature, but this same phenomenon is applicable in other applications that include a taboo feature (Law et al. 2007), and more widely, in other social computation applications such as collaborative filtering (Konstas, Stathopoulos et al. 2009), folksonomies (Golder and Huberman 2006; Schmitz et al. 2007), FlickR (Sigurbjörnsson and Zwol 2008), and other applications, where users see the words that were already contributed by previous players. In the following sections, we provide a theoretical basis for characterizing the possible biasing influence of taboo words, and we investigate it empirically. Practical implications are then discussed.

3 THEORETICAL DEVELOPMENT

3.1 Anchoring and Adjustment

Kahneman and Tversky (1974) introduced three heuristics that people use in everyday situations: representativeness, availability, and "adjustment and anchoring". The idea behind anchoring is that when asked to quantify something, people are influenced by reference points. A classic experiment asks subjects to estimate the percentage of African countries that are members of the United Nations, except that prior to that, half the subjects were asked to state whether the percentage was higher or lower than (say) 10% while another half of subjects was asked to state whether the percentage was higher or lower than (say) 65%. When they later gave their estimates of the actual percentage, those who had been asked higher/lower than 10% provided lower estimates than those who'd been asked higher/lower than 65%. Subjects did adjust their estimates upwards from the low anchor, and downwards from the high anchor, in the correct direction, but they didn't adjust enough, so that the anchor still had a considerable residual effect.

Anchoring effects have also been established for tasks that are not inherently numerical, and/or for which there is no right answer. For example, Cervone and Peak (1986) showed an anchoring effect with regard to subjects' self-efficacy, in a task with no right answer. Subjects in high (low) treatment groups were asked whether they could solve more or less than 18 (4, respectively) problems. When next asked how many they thought they could solve, subjects in the high-anchor group gave higher estimates. And, anchoring effects have been shown not only for strictly *numerical* questions, but also regarding other quantifiably *measureable* constructs. For example, Lebeof and Shafir (2006) review research in which subjects judged pleasantness of scents, musical pitch, and so on.

Drawing on anchoring theory, we predict that players in the ESP game will suggest words that are similar to the taboo words to which they are exposed. In the ESP game, users see taboo words, and are asked to suggest (other) words with the aim of matching their anonymous partner. We hypothesize that taboo words may serve as an anchor point. In this game, users must adjust off this anchor point according to the rules of the game, because the taboo words themselves are off limits. But the anchoring and adjustment heuristic would lead us to expect that the anchor will exert an influence even after the adjustment process, leaving the player to suggest words that are similar to the taboo. In effect, the labels that are generated by earlier player pairs set the tone for all subsequent label suggestions.

As an example, suppose players are suggesting labels for the image shown in Figure1. Assume that the word 'dog' is among the taboo words. This taboo word may exert a biasing influence that leads an average player to think to suggest words related to 'dog' – e.g. 'animal' -- as opposed to another word that he/she might otherwise have thought of.

However, to our knowledge the anchoring heuristic has only been previously considered for tasks that require the subject to suggest a quantity; in the ESP game, by contrast, he/she is asked to suggest words. It is difficult to advance theory-based arguments about the applicability of anchoring biases to this setting, because the psychological mechanisms behind the anchoring bias is not clearly known. Indeed, this has been the source of substantial criticism of this line of research (Mussweiler and Strack 1999). In order to advance the idea of an anchoring effect to word suggestions, we augment the anchoring and adjustment heuristic with ideas from priming theory.

3.2 Priming Theory

In order to provide the basis for our prediction that taboo words will have a biasing influence in the ESP game, we augment the anchoring perspective with the theory of priming, especially semantic priming. Semantic priming is a phenomenon whereby a stimulus such as a word 'primes' the mind to more readily respond to other, semantically related words. In a classic task called the lexical decision task (LDT), a user is supposed to decide, as quickly as possible, whether a set of letters (e.g. hospital) is a word or non-word; subjects are found to (correctly) respond more quickly when they had been previously primed through exposure to a semantically related word (e.g. nurse). In the experimental task that is most relevant to the ESP game, subjects in a free association task are presented with a word and are asked to respond with the first (other) word that comes to mind. Results show that subjects respond in predictable ways with semantically related words. In fact, a number of books have been formed to catalog the totality of known 'word association norms' (Palermo and Jenkins 1964).

Unlike the free association task, in the ESP game are not asked to name a word that comes to mind in response to the stimulus. On the contrary, they are (only) told *not* to suggest the taboo words themselves, and it is actually in the game owner's interest that they *not* specifically suggest semantically related terms.

3.3 Hypotheses Development

We have seen that taken individually, neither the anchoring heuristic nor priming theory directly predicts that ESP game players will be biased by taboo words. The anchoring heuristic has not been previously established for lexical tasks, while priming theory has been established for lexical tasks, but only where the subject is specifically asked for words that come to mind in response to the stimulus. We draw on the combination of these two established theories, to advance our prediction that when primed with taboo words, players who are asked to suggest different words will tend to suggest words that are semantically related to the taboo.

The idea of forging a combination of these theories is related to a stream of research that proposes that the theoretical mechanism behind anchoring is actually a kind of priming (Mussweiler and Strack 1999; Mussweiler 2002). They propose a two-step mechanism called selective priming. When a subject is first asked a relative question such as whether the percentage of African nations that are members of the UN is higher or lower than 10%, he/she generates in his/her mind evidence that is consistent with the hypothesis that the percentage is equal to 10%; this information is then primed and readily accessible in the person's mind, so that later, when asked for the actual percentage, he/she is influenced by the readily accessible information that points to something close to 10%. The first stage of their proposed mechanism is not exactly applicable here, because subjects are not given any initial task that requires them to think about the taboo words.

Based on the combination of ideas from anchoring and priming, and consistent with Mussweiler's work that the mechanism behind anchoring is actually a type of priming, we predict that players who are exposed to taboo words will go on to suggest words that are semantically related to the taboo words. The theoretical extension regards the scope of semantic priming: it is predicted to occur even when the person is not first tasked with generating ideas or words related to the stimulus (as he/she is in Mussweiler's work on anchoring) and even when he/she is not specifically asked what words come to mind in response to the stimulus (as in word association tasks from the priming literature). The theorized model is that exposure to the taboo words as words *to keep-away from*, activates semantically related words in the user's mind.

The nature of this extension of scope can be more sharply understood by comparison with the phenomenon of negative priming (Tipper and Driver 1988). When a subject is presented with a pair of stimuli and is instructed to attend to one (e.g. label it) and *ignore* the other, then it is found that subjects were less able (i.e. slower and less accurate) to identify a subsequently presented object from the same category. In other words, there was a semantic spreading of *ignoring*, a kind of opposite effect to the usual phenomenon of spreading activation.

Our setting – which is common to many social computation applications – involves a subtly different case. To ignore a stimulus, means to pay it no attention. In our setting, subjects are not instructed to ignore the taboo words, but to *avoid* them; to suggest a taboo word in the ESP game is to waste time, in a game for which time is limited. Avoiding previously-established words requires attending to them. Our theoretical extension to the work on semantic priming and anchoring is that *the kind of attention that is needed to avoid a word*, has the effect of spreading activation to semantically related concepts, which then are more readily retrieved.

> *H1: Players who are exposed to taboo words – words that they are to avoid -- will suggest words that are semantically related to the taboo words*

The implication of H1 is that when the ESP game is played by a population, the totality of labels that will be collected for a given image will be less varied as a result of the biasing effect. The reason is that at any given moment, all player pairs who are asked to provide labels for a given image see the same taboo words. If players are biased to suggest new words that resemble the taboo, it also means that the various players will be suggesting new words that resemble one another. The bottom line from the practical perspective is that the totality of labels that are collected for a given image will be less varied as a result of the biasing influence of taboo words. If we define that any word that any player pair matches on is accepted by the system as a label for that image, then we get the following:

> *H2: The totality of words that are matched-on for a given image will be less varied as a result of exposure to the taboo words.*

In the example of Figure 1 above, assume that 'dog' biases an average player to think of 'animal' (or other related term). Then we expect that a number of players may all suggest 'animal'. In this manner, when a taboo word exerts a biasing effect, the words that are suggested by different groups will be less varied, so there will also be less variety in the set of words on which different player pairs match.

To summarize, H1 is primarily focused on a theoretical phenomenon, and can shed light on the scope of the anchoring and priming effects in the context of social computing games where players may only contribute new labels, tags, etc. H2 is more focused on a practical implication of the predicted bias, according to which the variety – and therefore the value – of the games' informational by-product is adversely affected by this design choice.

4 EXPERIMENT

4.1 Treatment and Control Conditions

For the sake of experimental control and to make a null hypothesis for H1, we programmed two versions of an ESP game, one with visible taboo words and one with invisible taboo words. The idea of invisible taboo words is to retain all the same rules of the ESP game, including the feature that players can win only by matching on a non-taboo word, but without exposing users to the taboo words so that they will not bias users' word suggestions. Then, we compare the similarity of words that players suggest in the visible treatment, against those of the invisible treatment, expecting that players' exposure to the words in the visible treatment will result in biased word suggestions.

Invisible taboo words can be approached in a variety of ways. One way is to allow players to suggest any words including taboo words, except that the system simply ignores them in the sense that the system will not announce a winning match until the two partners submit a matching non-taboo word. An alternative, which we adopt, is to alert the player if and when he/she suggests a taboo word, that it is taboo. This would still mean that the user were exposed to the word's being taboo, but only after the fact, and only the one word he/she had suggested, and it would not continue to be displayed. – all factors that would logically mitigate the biasing influence of the taboo words, as compared with the usual practice of simply showing the full set of taboo words from the get-go. We opted for this version of invisible taboo words, over the first alternative in which users never see them, for two reasons: (1) It makes the games more nearly identical; (2) it is the more conservative approach. On the first point, if players were never shown that a word is taboo even after they suggest it, they might interpret their failure to match as evidence that their partner had not suggested that word, and this could influence their beliefs about what words his/her partner is inclined towards. Our approach, in which the system informs the player if he/she suggests a taboo word, makes the two games more nearly identical. On the second point, this is also the more conservative approach in terms of hypothesis testing. Rather than testing whether taboo words exert a biasing influence when players see them as compared to when they don't, we test whether they exert more influence when they are exposed to the taboo words in advance, as compared with when they are exposed to them only if and when they suggest one.

To summarize, H1 and H2 are both tested by comparing word similarities between the visible-taboo treatment and invisible-taboo treatment conditions. H1 tests whether the words that players suggest are more similar to the taboo words in the visible treatment than the invisible, and H2 tests whether the variety within the set of suggested and matching words is lower in the visible treatment than the invisible treatment.

4.2 Initial Taboo Word Seeds

In a natural setting, the system may start with no taboo words for a given image, or it may begin with words that are gleaned from other sources, such as keywords from the html "alt" field that a webpage author may use to label an image. In our laboratory setting, we artificially seeded the initial taboo words. This was done by asking a group of 4 research assistants to study the images and to supply a list of descriptive labels (any number of them), together with a weight for each label that reflects the degree to which they think the label fits the picture. The two or three words with the highest weights were chosen by the researchers as initial taboo words for the picture. As play of the actual ESP games progressed over many player pairs, any word that was matched by three (this threshold was chosen a priori, and based only on prior experience) player pairs was added to the list of taboo words for that image. The extended taboo word list would be in force for all subsequent player pairs when they encounter that image.

4.3 WordNet Similarity Measures

Our hypotheses predict that ESP game players will suggest words that are semantically related to the taboo words themselves. Testing this idea requires a measure of semantic relatedness. There are two approaches to measuring word similarity, i.e. symbolic and statistical. A statistical approach is based on word co-occurrences. An example of a statistical measure of word similarity is the Dice coefficient. However, a statistical approach is defined relative to a corpus, which in our setting is the entire corpus of all English text. We are not aware of any publicly available tool that returns a Dice (or similar statistical) similarity measure on the basis of the whole English language, or even a subset such as the Web. For this practical reason, we chose to use a similarity measure based on WordNet (Miller 1995), which is publicly available and has been used widely in research. Wordnet is a lexical database that defines word meanings, and organizes words into sets of synonyms called synsets, as well as a fixed set of relationships (e.g. is-a) among the synsets. A number of similarity measures have been advanced to measure the relatedness of any two Wordnet concepts (Pedersen, Patwardhan et al. 2004). We adopt the widely used Lin measure (Lin 1998), which has been shown to have a high correlation with human-assigned relatedness. We used a publicly available program called Wordnet::Similarity that implements this (and other) similarity measures (Pedersen et al. 2004).

To test H1, for each player working on a given image, we calculated the average proximity between each taboo word and each of the words that the player suggested. We did this for each picture and player, under each of the two treatment groups, visible and invisible taboo words. If H1 is correct that the taboo words bias the words that a player suggests for a picture, then the average proximity should be larger in the visible taboo words treatment as compared to the invisible taboo words treatment.

As an example, consider the picture and taboo words as shown in Figure 1. The screenshot shows that at that moment in time, there were two taboo words: dog, intelligent. Each of the guesses offered by the shown user – i.e. glasses, accountant, pencil – are compared with each of the taboo words. Let us suppose that his/her partner suggested one word, i.e. pencil, so that the pair matched on that label. This example, for a single image and players, is summarized in Table 1. Using Wordnet, we would calculate the proximity of word pairs as shown in Table 2.

Image #	**Taboo Words**	**Words suggested by player 1**	**Words suggested by player 2**
1	Dog Intelligent	glasses	pencil
		accountant	
		pencil	

Table 1. Example of taboo words and terms suggested by players

Image #	Player	**Visible Taboo Treatment**
1	1	Proximity(dog, glasses) = .16
1	1	Proximity(dog, accountant)=.2
1	1	Proximity(dog, pencil)=.16
1	1	Proximity(intelligent, glasses)=0
1	1	Proximity(intelligent, accountant)=0
1	1	Proximity(intelligent, pencil)=0
Avg image 1 player 1		.087
1	2	Proximity(dog, pencil)=.16
1	2	Proximity(intelligent, pencil)=0
Avg image 1 player 2		.08

Table 2. Example of calculations to test H1 based on data from Table 1

Tables 1-2 show only the similarity measures for a single pair of players on a single image, under one treatment condition, i.e. visible or invisible taboo. These calculations are repeated for all images and all players in the two treatment conditions, yielding two lists of numbers. With fifteen images and twenty players, each list had up to 300 numbers. The average of each list is calculated. The statistical test is a difference of means. Recall that taboo words for a given image may be added as player pairs matched on other terms, meaning that another player pair who saw the same image later might have faced (say) three taboo words: dog, intelligent, and (say) pencil. The calculations always consider the taboo words that were in force for that particular player pair.

The test for H2 is per-image. The idea is to consider the set of labels that is the informational by-product of the ESP game, in the form of a set of labels that emerge from the game for each image. For each image, we construct a set of labels that is the union of the initial taboo words and the labels on which at least one pair of players matched. We calculate the pairwise proximity between every pair of words in this set, except for pairs of taboo words which are equivalent in the two treatment conditions and so do not affect our test statistic. We do the same for every other image, and test the difference in means between the visible and invisible treatment conditions. The idea is to capture the variety (opposite of similarity) among all the words that emerge as labels for each image. Table 3 shows an example of these calculations for a single image, assuming that there were two initial taboo words, i.e. dog and intelligent, and that in the visible treatment at least one group matched on animal and one on smart (and no player pair matched on any other label), while in the invisible treatment at least one player pair matched on funny, and at least one on puppy, and at least one on glasses. The average of these proximities is calculated for each image. With fifteen images, there are fifteen numbers for each treatment condition.

Visible Taboo Treatment	**Invisible Taboo Treatment**
Proximity(dog, animal) = .76	Proximity(dog, funny)=0
Proximity(dog, smart)=0	Proximity(dog, puppy)=.83
Proximity(intelligent, animal)=0	Proximity(dog, glasses)=.16
Proximity(intelligent, smart)=0	Proximity(intelligent, funny)=0
Proximity(animal, smart)=0	Proximity(intelligent, puppy)=0
	Proximity(intelligent, glasses)=.08
	Proximity(funny, puppy)=0
	Proximity(funny, glasses)=0
	Proximity(puppy, glasses)=.14
Avg() for image 1 = .15	**Avg for image 1** =.13

Table 3. Example of calculations to test H2, based on a single image

4.4 Experimental Sessions

Forty volunteers were recruited, primarily from a high-tech company and a technical high school. Players were randomly assigned to one of two treatment conditions, visible and invisible taboo words. The usual ESP game instructions were given. It was explained that each player would be paired with an anonymous partner who was playing simultaneously. The players were told they would see a series of pictures, and that for each one, they should suggest words until they and their partner both suggested the same word (among others), at which point they would receive points for that picture and move on to the next. Players were also told about the existence of taboo words that could not be the basis for a match, and in the invisible version of the game, it was further explained that the system would alert them if they tried to suggest such a word. The full game was limited to two and a half minutes or fifteen pictures, whichever came first. A player could decide to pass on a given image at any time; this would mean that he/she and his/her partner would move on to the next picture even if they had not yet matched on any label for the current picture. In each treatment condition, half the player pairs received the images in one order, and half in the opposite order.

4.5 Computerized Game

A client-server application was programmed in the C# language in a .net environment. The program is TCP-based but not Web-based, and supports bi-directional communication, i.e. client to server and vice versa, which is more robust than the more usual request-reply architecture.

4.6 Results

Because there were 20 players in each condition and a maximum of fifteen pictures, there was a possible maximum of 300 instances of image-players in each treatment, or 600 in total. In practice, there were actually 572 word pairs; 28 times the player didn't submit any word with a valid semantic proximity to either of the taboo words.

The test of H1 showed that the average similarity of matching words to taboo words was .21 in the visible taboo treatment and .18 in the invisible taboo treatment. This difference supports H1 ($t(570) = 3.777$, $p < .001$).

The test of H2 showed that the average similarity among all pairs of words that emerged as the labels for each image was higher in the visible taboo treatment (mean = .24) than the invisible taboo treatment (mean = .17). This difference supports H2 ($t(28) = 2.19$, $p < .05$).

4.7 Discussion

Consistent with our theoretical predictions, players in the visible-taboo treatment tended to match on words that are related to the taboo words, and this lessened the variety of words that ultimately were generated by the game for each image. As noted at the outset, this means that the taboo words have a self-limiting effect; yes, players are forced to suggest words other than the taboo words themselves, but they are paradoxically influenced to suggest terms that are similar to the taboo. We think of this as a "don't look here" effect.

There exists an alternative explanation of this behaviour. It is possible that a player may not be biased him/herself, but may believe that his/her partner may be, and so it makes sense to suggest related terms in the hopes that that's what the partner will do. Or, to continue this line of thinking, it is possible that players simply use the information in the taboo words to help them coordinate. That is, neither player is necessarily biased, but they purposely suggest terms near the taboo words in the hopes that the partner will do the same. This sort of behaviour would be in line with the literature on focal points (Mehta, Starmer et al. 1994). This alternative explanation does not diminish the practical effect, and the two explanations are not exclusive of each other, but further work is needed to disambiguate the two theoretical mechanisms through which taboo words exert their effect.

5 CONCLUSIONS AND CONTRIBUTIONS

The premise of the social computing and human computation paradigms is that human interaction performs computations, and its informational by-product can mimic the output of machine computations. It is perhaps no great surprise that this analogy is imperfect, and the information that is produced is influenced by the type of computation. Our work demonstrates one such effect, namely, that people are influenced by what previous players have contributed.

Our research contributes to the growing literature on social computing by illuminating a specific imperfection that is found when the results of previous participants are shown to subsequent participants. In the specific case we studied empirically, the work of previous participants takes the form of “taboo” words. In other social computation applications, such as FlickR, the feature is not called “taboo”, but the effect is the same, i.e. previously-assigned tags are visible to the user, who cannot assign them again. We found that subsequent users are influenced to suggest words that are related to the taboo. The result is that the taboo feature, which is intended to ensure term variety, is self-limiting. A possible – if partial -- remedy is the invisible taboo feature, which we had used as a methodological device, but which can also be implemented as a game feature to limit the biasing effect of taboo words. Internet businesses (e.g. Digg, Flickr) that depend on tags may see fit to alter their interface designs to limit the biasing effects shown here.

Our work also contributes to the literature on priming and anchoring. Anchoring theory has been previously applied only to tasks in which subjects quantify some variable, not to tasks in which they suggest a word. Regarding priming, positive priming effects have previously been shown for word tasks in which the task was specifically to suggest a word that comes to mind in reference to a stimulus, which is not our case; and negative priming effects have been previously shown for tasks in which the subject is instructed to *ignore* a stimulus, which is also not our case. Our work extends the known scope of priming phenomena, by demonstrating a *positive* priming effect when the subject is exposed to a stimulus and instructed to *avoid* it. This result may also have further theoretical implications, in terms of disambiguating between various hypothesized mechanisms behind priming effects; pursuing these implications is left for future work.

References

Cervone, D. and P. K. Peake (1986). Anchoring, Efficacy, and Action: The Influence of Judgmental Heuristics on Self-Efficacy Judgments and Behavior. Journal of Personality and Social Psychology, 50(3): 492-501.

Chan, K. T., I. King and M.-C. Yuen (2009). Mathematical modeling of social games. In *Proceedings of Computational Science and Engineering, 2009 (CSE'09)*, IEEE, Vancouver

Dryer, D. C., C. Eisbach and W. S. Ark (1999). At what cost pervasive? A social computing view of mobile computing systems. IBM Systems Journal, 38(4): 652-676.

Goh, D. H.-L., R. P. Ang, C. S. Lee and A. Y. K. Chua (2011). Fight or Unite: Investigating Game Genres for Image Tagging. Journal of the American Society for Information Science and Technology, 62(7): 1311-1324.

Golder, S. A. and B. A. Huberman (2006). The Structure of Collaborative Tagging Systems. Journal of Information Science, 32(2).

Konstas, I., V. Stathopoulos and J. M. Jose (2009). On Social Networks and Collaborative Recommendation. In *Proceedings of 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*, ACM, Boston

Law, E. L. M., L. v. Ahn, R. B. Dannenberg and M. Crawford (2007). Tagatune: A game for music and sound annotation. In *Proceedings of The Eighth International Conference on Music Information Retrieval*, Austrian Computer Society, Vienna, Austria

LeBoeuf, R. A. and E. Shafir (2006). The Long and Short of It: Physical Anchoring Effects. Journal of Behavioral Decision Making, 19: 393-406.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of The Fifteenth International Conference on Machine Learning (ICML-98)*, Madison

Mehta, J., C. Starmer and R. Sugden (1994). The Nature of Salience: An Experimental Investigation of Pure Coordination Games. The American Economic Review, 84(3): 658-673.

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11): 39-41.

Mussweiler, T. (2002). The Malleability of Anchoring Effects. Experimental Psychology, 49(1): 67-72.

Mussweiler, T. and F. Strack (1999). Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. Journal of Experimental Social Psychology, 35: 136-164.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, Muthén & Muthén.

Palermo, D. S. and J. J. Jenkins (1964). *Word Association norms : grade school through College*. Minneapolis, University of Minnesota Press.

Pedersen, T., S. Patwardhan and J. Michelizzi (2004). WordNet::Similarity: measuring the relatedness of concepts. In *Proceedings of HLT-NAAC*, Association for Computational Linguistics, Stroudsburg

Quinn, A. J. and B. B. Bederson (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of CHI 2011*, Vancuover

Schmitz, C., M. Grahl, A. Hotho, G. Stumme, C. Cattuto, A. Baldassarri, V. Loreto and V. D. P. Servedio (2007). Network Properties of Folksonomies. In *Proceedings of WWW 2007,* Banff

Sigurbjörnsson, B. and R. v. Zwol (2008). Flickr Tag Recommendation based on Collective Knowledge. In *Proceedings of WWW 2008*, ACM, Beijing

Tipper, S. P. and J. Driver (1988). Negative priming between pictures and words in a selective attention task: Evidence for semantic processing of ignored stimuli. Memory and Cognition, 16(1): 64-70.

Tversky, A. and D. Kahneman (1974). Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157): 1124-1131.

von Ahn, L. (2006). Games With a Purpose. IEEE Computer, June: 96-98.

von Ahn, L. and L. Dabbish (2004). Labeling Images with a Computer Game. In *Proceedings of CHI 2004,* ACM, Vienna, Austria

Von Ahn, L. and L. Dabbish (2008). Designing games with a purpose. Communications of the ACM, 51(8): 58-67.

Von Ahn, L., M. Kedia and M. Blum (2006). Verbosity: a game for collecting common-sense facts. In *Proceedings of SIGCHI conference on Human Factors in computing systems (CHI 2006),* ACM, Montreal

Von Ahn, L., B. Maurer, C. McMillen, D. Abraham and M. Blum (2008). reCAPTCHA: Human-based character recognition via web security measures(2008): 1465-1468. Science, 321(5895): 1465-1468.