# Collabio: A Game for Annotating People within Social Networks

**Michael Bernstein[1], Desney Tan[2], Greg Smith[2], Mary Czerwinski[2], Eric Horvitz[2]**

[1] MIT CSAIL
32 Vassar Street, Cambridge MA 02139
msbernst@mit.edu

[2] Microsoft Research
One Microsoft Way, Redmond WA 98052
desney | gregsmi | marycz | horvitz@microsoft.com

## ABSTRACT

We present Collabio, a social tagging game within an online social network that encourages friends to tag one another. Collabio's approach of incentivizing members of the social network to generate information about each other produces personalizing information about its users. We report usage log analysis, survey data, and a rating exercise demonstrating that Collabio tags are accurate and augment information that could have been scraped online.

### Author Keywords

Social computing, social tagging, human computation.

### ACM Classification Keywords

H5.2. Information Interfaces: User Interfaces.

## INTRODUCTION

Crowdsourced applications depend critically on having a broad enough appeal to attract a community of contributors. However, it can be difficult for such systems to succeed when there is only a small network qualified to provide the information: for example, when gathering information about a particular individual. Two challenges arise in such a situation: motivating enough members of the small network to participate, and ensuring the accuracy of the generated information. To address these problems, our work explores mechanisms for gathering social information in a social context. We specifically take on the challenge of personalization by building a social tagging game. The game adapts elements of Games with A Purpose [6] and extends their design principles using social controls. While previous work has explored people tagging [2], broadening this beyond the enterprise creates challenges around motivation, tone, and accountability that we address in our work.

In this paper, we present Collabio (*Collab*orative *Bio*graphy), a game that elicits descriptive tags for individuals within the Facebook social network. Collabio (Figure 1), available at http://apps.facebook.com/collabio, elicits information that your friends know about you. This information includes your personality, expertise, artistic and musical tastes, topics of importance, and quirky habits. The application leverages properties of the social network such as competition and social accountability to solve the tag motivation and accuracy problems within a social framework.


*UIST'09*, October 4–7, 2008, Victoria, BC, Canada.


## RELATED WORK

Fringe Tagging is a social people-tagging application developed for internal use at IBM [2], and Collabio extends Fringe's exploration. Beyond the differences inherent in an enterprise application, Fringe takes a largely pragmatic perspective on motivating people to participate: it "enables people to organize their contacts into groups, annotate them with terms supporting future recall, and search for people by topic area" [2]. Collabio, by contrast, is oriented primarily toward encouraging social connectedness, and is a game.

Research systems such as Mr. Taggy [4], Spar.tag.us [3] and Dogear [5] explore the social tagging of web content. Collabio adds to the knowledge of these systems because its motivation is derived from "tagging for you" rather than "tagging for me" (Dogear, Fringe) or "tagging for us" (Spar.tag.us). Unlike in the other tools, Collabio taggers do not receive direct benefit by being active; instead, they hope to incentivize their friends to reciprocate and tag them.


**Figure 1. The user has guessed several tags for John Smith, including *band, ohio* and *vegas*. Tags guessed by John's other friends are hidden by dots until the user guesses them.**

Collabio also builds on existing social people-tagging applications on Facebook. These applications typically aim to maximize entertainment rather than quality of tags. iDescribe[1] and Compare People[2] allow users to place predefined tags on their friends. This assumes a small set of static tags and does not leverage the richness of knowledge in the network. Describe Me[3], Define Me[4], and Impressions[5] encourage users to author new tags. However, they allow authors to see and reapply existing tags, hence potentially skewing perception and reducing the actual richness of tags.

Our approach is inspired by Human Computation [6], which aims to obtain useful information for computers by enticing users to provide it. We extend the design principles of these games: specifically, though Collabio utilizes game motivations such as point scores and leader boards, it leans just as heavily on social motivators such as the practice of returning positive or negative actions in kind. Rather than anonymously pairing random players to prevent cheating, we explicitly target users within established social groups to contribute data, relying on social accountability and self-interest in maintaining a positive reflection of yourself in your profile. Finally, rather than gathering information common to all web-enabled humans, we directly target information that is known and verifiable only by a small social group: information about a friend. IBM's Dogear social bookmarking game shares some of these characteristics; Collabio focuses the information on a single individual at a time and makes the collected information visible in anonymized form to the rest of the social network [1].

## COLLABIO

Collabio is embedded in the Facebook social network. In the following sections, we describe Collabio's three top level interface tabs: the tab in which users can *Tag!* their friends, the one in which they can manage *My Tags*, and the one in which they can see the *Leaderboard*. We then discuss propagation through the social network, the incentive design space, and issues of cheating and abuse.

### *Tag!* Friends

The main activity of Collabio is tagging friends, so the focus of the user's experience is the tagging page (Figure 1). Players see the tag cloud that other friends have created by tagging the selected friend. When presenting this cloud, Collabio only displays tags that the user has already explicitly guessed. Tags not yet guessed are obscured by replacing each constituent letter with a solid circle; for example, the tag *UIST* appears as ●●●●. Whitespace in obscured tags is represented by clear circles such as ○. Thus, the length and makeup of the obscured tag provide hints as to the hidden text. As an additional hint, terms in the tag cloud are alphabetically ordered. The tags in the cloud are scaled so that the popular tags are larger.

As the user tags a friend, one of two things happens. If the tag is new and has not previously been placed on their friend, the tag is inserted into the cloud. If the tag exists, then it is revealed within the cloud. For each guess, users receive points equal to the total number of people who have applied a tag, including themselves. If they are the only person to have guessed that tag, then they get 1 point; if there are 11 others, they get 12 points. These points continue to accumulate as more people apply the tag, so earlier taggers' scores rise as well. A user can retract a tag by clicking on a small × by the tag.

To expose one's score to others, and to stimulate competition, each tagged friend has a "People who know [this friend] best" pane which lists friends who have earned the largest number of points from tagging that friend (Figure 1).

In the current system, if the user is the first to tag a friend, Collabio seeds the tag cloud with terms from the friend's public profile (such as network names, affiliations, or interests), ensuring that the tag cloud is never completely empty. These tags are attributed to the Collabio Bot. We observed early on that users were typically unwilling to tag others who had not already added the application, so this tag seeding is helpful in overcoming reluctance to be the first to tag.

### Managing *My Tags*

The *My Tags* interface allows users to inspect and manage tags their friends have placed on them. The My Tags page contains three sections: a fully uncovered tag cloud, a full top scorers list allowing the user to see every friend who has tagged them, and a table explaining which friends tagged the user with which tags. In order to allow people to maintain control of tags placed on them, Collabio allows them to easily delete tags from their tag cloud by clicking on a small × by the tag.

### The *Leaderboard*

The third Collabio tab is the Leaderboard. While the individual leaderboards on the *Tag!* tab encourage users to keep tagging a friend until they are listed as one of the Top Friends for that person, the global leaderboards encourage users to continue tagging activity within the application. We present two lists here, one of the friends that have the most unique tags placed on them, and the other of the friends who have tagged the most other friends.

### Designing for Social Spread

Collabio relies on social mechanisms to spread to new users and retain existing ones. For example, the individual leaderboards are labeled "friends who know [this friend] best" to conflate closeness of friendship with score in the game. More subtly, when a user tags a friend, the friend receives a Facebook notification whether or not that friend has previously played Collabio:

Michael Bernstein has tagged you with **cyclist** and 7 other tags using Collabio. Tag Michael back, or see what you've been tagged with. 2:41pm

[1] http://apps.facebook.com/idescribe
[2] http://apps.facebook.com/comparepeople
[3] http://apps.facebook.com/describeme
[4] http://apps.facebook.com/defineme
[5] http://apps.facebook.com/impression

A similar version appears on the tagger's wall feed and on the homepage Facebook news feed.

### Dealing with Cheating and Abuse

Because Collabio activity can only occur between people with a mutually-established social connection, we rely on social pressures to prevent cheating. Possible cheating mechanisms include asking friends for the answers, reverse engineering tags using a search strategy on the alphabetized cloud, using a large number of nonsensical tags (e.g., *a*, *aa*, *aaa*), and tagging an individual with an undesirable tag as a joke or punishment. Some of these strategies are more work than guessing tags themselves, or result in suboptimal scoring performance. Others dirty the tagged individual's tag cloud, and generally result in the tagged individual becoming annoyed at the tagger or retracting points.

## IMPLEMENTATION

The Collabio application interface is built as an AJAX-enabled ASP.NET web application, which calls a SQL Server-backed Windows Communication Foundation web service for data storage and querying.

## FIELD DEPLOYMENT AND EVALUATION

We analyzed tag statistics collected from July 2008–March 2009. In that time, Collabio gathered 7,780 unique tags on 3,831 individuals in 29,307 tagging events. These tags were generated by 825 different users out of 1,350 who installed the application according to Facebook. The median user who tagged at least one friend received 11 unique tags in return, indicating that even minimal usage of Collabio resulted in a user being relatively well-tagged by friends. Being well-tagged is a positive outcome because it positively reinforces playing time and it produces data to begin applying to the personalization problem.

### User Survey: Tag Accuracy and Popularity

We supplemented our log data with a survey aimed at active Collabio users, a group we defined as users who had tagged at least three friends, were tagged by at least three friends, and had at least nine distinct tags. Using Facebook's notification service, we invited 112 of Collabio's active users to fill out a survey about their experience. Forty-nine users (24 female) responded to the survey. The median age was 27 ($\sigma$ = 4.1), and respondents seemed slightly skewed toward students and researchers with an interest in user interfaces. We offered a small gratuity for responding.

To learn more about tag content, we asked each survey respondent to rate nine tags in their tag cloud. These tags were drawn from three buckets: *Popular Tags*, the three tags applied by the most friends; *Middling Tags*, drawn randomly from tags that occurred at least twice but less often than the Popular Tags; and *Unique Tags*, drawn randomly from tags applied by only a single friend. For each tag, the user provided a rating on a 7-point Likert scale (1 for disagreement and 7 for agreement) for each of two questions: "This is a good tag for me," and "This tag is something I would expect lots of people to know about me." In addition, participants classified each tag into a type category.

| | Popular Tags | Middling Tags | Uncommon Tags |
|---|---|---|---|
| Accurate | $\mu = 6.42$ $\sigma = 0.92$ | $\mu = 5.83$ $\sigma = 1.39$ | $\mu = 5.13$ $\sigma = 1.61$ |
| Widely known | $\mu = 6.22$ $\sigma = 1.22$ | $\mu = 5.21$ $\sigma = 1.58$ | $\mu = 4.14$ $\sigma = 1.77$ |

**Table 1. User ratings of how accurate and widely known the tag buckets were, from 1 (very inaccurate / not widely known) to 7 (very accurate / widely known).**

*Results*

Respondents indicated that a large percentage of Collabio's tags comprise affiliations, interests, expertise and hobbies (e.g., *MIT*, *Atlanta*, *HCI*, *tennis*); the long tail of tags contribute a wide variety of more unusual information. Popular Tags were reported to be mainly affiliations; Middling Tags and Uncommon Tags were more commonly reported to capture interests, expertise and hobbies. The Uncommon Tags also captured several unusual pieces of information categorized by respondents as 'miscellaneous', including clothing choices, special abilities, and a friend's dog.

Generally, the more popular the tag, the more accurate it was and the more well-known the fact. Survey participants rated all three classes of tags as accurate descriptors of themselves, and all but Uncommon Tags as known by many people (Table 1). We ran a one-way ANOVA and found significant effects of tag bucket on goodness of tag ($F_{2,384}$=34.5, $p<0.001$, $\eta^2$=.15) and expectation that others know the given fact ($F_{2,384}$=67.1, $p<0.001$, $\eta^2$=.26). Pairwise posthoc comparisons using Bonferonni correction confirmed all factor levels were significantly different from each other in terms of accuracy and anticipated popularity.

We were surprised to find that even the Uncommon Tags were rated as accurate descriptors, and preliminary observations suggest that there is not much inaccurate information in the Collabio tag database. This result runs counter to our expectation, drawn from the literature [6], that repetition and independent agreement are important to guarantee accuracy. We hypothesize that social motivators were powerful enough in Collabio to avoid serious misuse or off-topic tags, but confirming this hypothesis remains future work.

### Rating Exercise: Tag Novelty

Our results suggest that Collabio generates accurate tags that are reasonably ordered by importance. However, if these tags are available elsewhere, we have not significantly advanced the state of the art. Could an algorithm or individual outside the social network create these tags by mining information available in users' Facebook profiles or the web and reproduce the relative ordering of tags?

To answer this question, we conducted a rating exercise. We recruited four native English speakers comfortable with Facebook and web search, but who had never used Collabio and did not know any Collabio users, to serve as raters. We tested whether our human raters, as a reasonable upper bound on machine inference, could find the tags on the Collabio users' profiles. Raters judged the set of tags under two scenarios: first using only the individual's Facebook profile available to friends, and second using only web search.

We randomly selected twenty survey respondents from the forty-nine who completed our previous survey as target individuals for our raters to investigate. We utilized the nine tags each target had rated in the survey, as well as three randomly-selected *Fake Tags* that were false and thus should not appear anywhere associated with the individual.

For each target individual, raters were presented with the twelve tags in random order and asked to rate each on a 7-point Likert scale according to the following statement: "I can find strong evidence that the tag applies to this individual." Raters were trained to give a score of 7 if the tag appeared verbatim, a score of 1 if there was no evidence in support of the tag, and a score of 4 if moderate inference was required based on the available evidence (e.g., the tag was *Atlanta* but the only relevant evidence was that the person attended Georgia Tech); the other values on the ordinal scale captured in-betweens. Raters were trained on example tags and profile sets until satisfactory agreement on the scoring scale was achieved. We randomized the order in which raters viewed individuals.

We also wanted to investigate whether our raters could determine how popular a tag had been, as verified by our survey data. For each individual, raters placed each tag into its original bucket: Popular, Middling, Unpopular, or Fake. Raters were told that three tags came from each bucket.

### Results

Raters evaluated tag evidence on Facebook and the web for a total of 480 tags across the twenty individuals. Cronbach's alpha was calculated to measure agreement between the raters, producing an overall agreement score of .82.

Experts found more supporting evidence for the more popular tag buckets, both on Facebook and the web (Table 2). A two-factor ANOVA comparing the effect of tag bucket (Popular vs. Middling vs. Uncommon vs. Fake) and evidence type (Facebook vs. Web) on rating found a main effect of tag bucket ($F_{3,1915}$ = 270.0, $p<0.001$, $\eta^2$=.30), and pairwise Tukey posthoc comparisons (all significant $p<0.001$) suggest that the more popular a tag was, the higher rating it received and so the easier it was to find evidence for. Thus, the more popular the tag, the more likely it occurred in a publicly visible area. We found no main effect of Evidence type, and inspection suggests that the scores between Facebook and the web are nearly identical.

In the bucket identification task, raters were the most reliable at identifying the extreme buckets: Popular Tags and Fake Tags. Raters had the poorest performance on Middling Tags and Uncommon Tags, correctly recognizing only about 40% of each. Thus, beyond the most common tags, it was difficult for non-friends to reconstruct tag rankings.

Overall, raters found evidence supporting Popular Tags, but moderate inference was required for Middling Tags and very little evidence was available for Uncommon Tags. Our original survey respondents indicated that even Uncommon Tags were generally accurate, so we may conclude that Collabio is collecting accurate information with Middling and Uncommon Tags that would otherwise be difficult or impossible to acquire, at least with simple online scraping techniques. This information comprises a majority of Collabio tags, since the Popular tags are by comparison few in number. Raters had considerable difficulty distinguishing Middling from Uncommon tags, and Uncommon from Fake Tags, so beyond the most obvious information it may also be difficult for a human, and certainly a machine, to recreate Collabio's tag ordering even coarsely.

| | Popular Tags | Middling Tags | Uncommon Tags | Fake Tags |
|---|---|---|---|---|
| Facebook Evidence | $\mu$ = 5.54 $\sigma$ = 2.36 | $\mu$ = 4.20 $\sigma$ = 2.68 | $\mu$ = 2.87 $\sigma$ = 2.56 | $\mu$ = 1.56 $\sigma$ = 1.76 |
| Web Evidence | $\mu$ = 5.72 $\sigma$ = 2.29 | $\mu$ = 4.17 $\sigma$ = 2.81 | $\mu$ = 3.04 $\sigma$ = 2.65 | $\mu$ = 1.5 $\sigma$ = 1.4 |

**Table 2. Mean ratings applied to tags, from 1 (no evidence to support tag) to 7 (tag appeared verbatim).**

## CONCLUSION

We have presented Collabio, a social network application that extracts latent personalizing information by encouraging friends to tag each other with descriptive terms in a game. Collabio has been successful in motivating players to tag over 3,800 people with tags that are both accurate and contain information not available elsewhere. We have further reported on usage log analysis, survey data, and a rating exercise showing that Collabio tags are relatively accurate and may add information to traditional methods of collecting tags. Documenting the design decisions and examining them in more detail remains future work.

## ACKNOWLEDGMENTS

We thank the MSR VIBE and ASI groups, as well as the 2008 interns, for helping us improve and spread Collabio. Many thanks also to Rob Miller and David Karger at MIT for continued feedback and advice on this work.

## REFERENCES

1. Dugan, C., Muller, M., Millen, D. R., et al. 2007. The Dogear Game: A Social Bookmark Recommender System. In *Proc. GROUP '07*, 387-390.
2. Farrell, S., Lau, T., Nusser, S., et al. 2007. Socially Augmenting Employee Profiles with People-Tagging. In *Proc. UIST '07*, 91-100.
3. Hong, L. & Chi, E. H. 2009. Annotate once, appear anywhere: collective foraging for snippets of interest using paragraph fingerprinting. In *Proc. CHI '09*, 1791-1794.
4. Kammerer, Y., Nairn, R., Pirolli, P., & Chi, E.H. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. CHI '09*, 625-634.
5. Millen, D., Feinberg, J., & Kerr, B. 2006. Dogear: Social bookmarking in the enterprise. In *Proc. CHI '06*, 111-120.
6. von Ahn, L. & Dabbish, L. 2004. Labeling Images with a Computer Game. In *Proc. CHI '04*, 319-326.