The Decreasing Marginal Value of Evaluation Network Size

Zheng Dong  
School of Informatics and Computing  
Indiana University  
Bloomington, IN 47408  
zhdong@indiana.edu

L. Jean Camp  
School of Informatics and Computing  
Indiana University  
Bloomington, IN 47408  
ljcamp@indiana.edu

Abstract  
The best way to protect information is never to release it. Yet even the earliest definition of security recognizes availability as a necessary quality. In this work, we seek to quantify the value of information disclosure for web resource evaluation and discovery. Communal evaluation tools help users share ratings on websites, music, and other online resources. This approach assumes that experiences are self-similar, so that a site one person visits is likely to have been evaluated and thus visited by others. Collaborative search tools aim for discovery as opposed to evaluation. Therefore, they assume participants in a collaborative network have large sets of non-overlapping sites so that an increase in network size corresponds to an increase in web coverage. We quantify the value of information sharing for these closely related but sometimes distinct functions.

In this paper, we analyzed a dataset that includes eight weeks of browsing history of 1084 college students that live in the same dormitory. Our experiments showed that for discovery, more sharing monotonically improves results; for evaluation, however, there are decreasing marginal returns for each participant added to the network. The subject population was selected for its homogeneity in order to mimic a collaborative network.

Categories and Subject Descriptors  
K.4.1 [Public Policy Issues]: Privacy; K.6.m [Miscellaneous]: Security

General Terms  
Security, Empirical Analysis

---

6 This material is based in part upon work supported by the National Science Foundation under Grant Numbers CNS 0943382 and IIS 1036918. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Keywords
Network size, collaborative search, communal evaluation

Introduction
Metcalfe’s Law (Metcalfe 1995) argues that the value of a network increases proportionally to the square of the number of participants. This implies there is always an increase in value as people join a collaborative network. Yet in terms of security and trust, the activity of spammers, phishers and malware distributors makes the converse case. In this work, we examine the value of network size in two domains: discovery and evaluation. We find that for discovery (specifically in collaborative search), more participants improve coverage, while for evaluation (specifically reputation systems for websites), a small number of participants are sufficient for a high degree of coverage. This is complementary to the observation that evaluation requires more trust than simple discovery.

Collaborative web services utilize browsing histories for two primary purposes: collaborative search and communal evaluation. For collaborative search, the main purpose is discovery. We expect sharing for discovery to result in a broader view of the web. Each participant ideally contributes new information. Researchers have developed tools that enable domain specialists (e.g. attorneys and service engineers) use search histories collaboratively (Komlodi & Lutters 2008). Experiments showed that these schemes help experienced employees annotate, communicate, and share more efficiently.

Evaluation requires some degree of homogeneity so that suggestions or ratings made by one are useful to others. This idea has been widely implemented in many collaborative filtering or ranking systems. For example, in Net Trust (Camp 2006), Delicious™, and Zigtag™, participants are able to access their own history and opt to share their history with a self-selected collaborative network. G+ has enabled integration of search results and social networks. Ratings for visited sites are recorded, and each member in the group can better identify malicious sites by using implicit and explicit reviews from other participants. In addition to site ratings, web services have been created to collect preference information anonymously from other similar participants, which filter certain content, and therefore provide an improved recommendation for music (Cohen & Fan 2000), movies (Park & Pennock 2007), etc.

However, collaboration among a group of participants brings forward issues of privacy and trust. For example, researchers have found that incidental information can be exposed during a co-located collaboration (Hawkey & Inkpen 2005). That means one’s private information may be viewed by another user in the same ad-hoc network. In addition, researchers pointed out that popular social networking websites may release users’ private information in the aggregation of their data (Chew 2008). AOL made the same

---

7 In this paper, we refer to the individuals who are browsing as “participants” in a collaborative network as opposed to “users” of computers.
8 http://www.delicious.com
9 http://www.zigtag.com
point inadvertently (Arrington 2006). Approaches have been proposed providing privacy-preserving mechanisms in collaboration. One approach to designing for trust would be to begin with the question, “How many participants do we need for effective collaboration?” The data here illustrate that more participants may not be better, depending on the application domain and collaborative goal.

On the one hand, a certain amount of shared browsing information is clearly useful. Nearly every collaborative network is improved by increasing the number of participants. The more information received from other participants, potentially the more accurate the result. On the other hand, more participants inherently increase information disclosure regardless of the benefits. The more people with whom information is shared, the more likely that personal information is released. Smaller groups may also have stronger trust relationships, making them more resilient to Sybil attacks and infiltration. Therefore, we need a balance between privacy and information sharing. In this work, we try to quantify that tradeoff in the domains of identification of untrustworthy sites and collaborative search.

In this paper, we analyzed a dataset that includes eight weeks of browsing history of 1084 college students that live in the same dormitory. Our experiments showed that for discovery, more sharing monotonically results in an improved discovery; for evaluation, however, there are decreasing marginal returns for each participant added to the network. The subject population was selected for its homogeneity in order to mimic a social network. The main contributions of this work are twofold. First, we have shown that for collaborative search, more participants enable more accurate results; while in communal evaluation, only a small number of participants are sufficient. That is, we have illustrated that the small collaborative networks can be highly effective and privacy-preserving. Because of the homogeneity of the population, it is likely that we have understated our results. That is, in a more heterogeneous community, the increase in search space (i.e. union) is likely to be greater while the increase in evaluation space (i.e. intersection) is likely to be less.

Our second contribution is a lesser methodical one. We provided two methods for the basis of our claims, sites vs. clicks, and performed an empirical analysis for these two challenges: discovery or evaluation. In addition, we proposed three ways of considering URLs: domain name, host, and first-level directory. This provides insight for the participant activity analysis. The most fundamental contribution is the insight on the value of additional collaborators as network size increases data sharing, risks of data leakage and risks of malicious participants.

The rest of the paper is organized as follows. Section 2 introduces the related work on communal evaluation, collaborative search, and privacy concerns in collaboration. Section 3 elaborates the experimental design. Section 4 reports and analyzes the results. Section 5 summarizes our findings, and concludes the paper.
Related Work

Reputation Systems for Web Resources
Malicious websites are proliferating with the expansion of the web. The APWG (APWG 2010) reported that the number of unique phishing websites detected on the second half of 2010 was up to 42,624. Since it is impossible for a single participant to know all malicious websites, researchers have been working on mechanisms that allow groups to share their browsing histories.

Communal evaluation has been widely used in web services. Efficacy of crowd sourcing vs. centralized systems in identification of malicious websites is subject to debate (Moore & Clayton 2008). Net Trust (Camp 2006) is a collaborative websites reputation system that collects ratings from participants within the same social network. Net Trust records explicit and implicit (e.g. history) ratings from participants, and shows the scores and suggestions in a browser toolbar. By this means, fraudulent websites, though they may appear the same as the real ones, can be better identified. Electronic commerce websites, such as Amazon and eBay rank their products and generate recommendations from activities of previous customers. Collaborative evaluation of URLs in spam has also been proposed (Thomas et al. 2011).

Social news websites such as Digg\textsuperscript{10} aggregate ratings for news stories, such that only the most popular stories appear on the first page. By visiting Digg, news subscribers have the opportunity to only focus on the most viewed news stories, which are subject only to positive and negative ratings. Zigtag is a social bookmarking website that enables their users to include tags on their saved web pages but not rate. Tags can be viewed not only by the creator, but also friends in the same social network. The Yahoo! social bookmarking website Delicious also enables tagging but not rating. In all cases, pages with the most tags are presented on the main page of the website.

A method for re-ranking movies (Park & Pennock 2007) was introduced by Park et al., which not only leveraged the traditional recommendation systems, but also integrated results from search engines. Markins et al. proposed a website to share and analyze bookmarks of participants in a social network (Markines, Stoilova & Menczer 2006). The system provides both recommendation and search functionalities. We selected for this work a reputation system that did not integrate search. We want to compare reputation for an identified object as opposed to recommendation, which includes search.

Komlodi et al. surveyed experienced workers such as attorneys and service engineers, and both groups showed high demand for a collaborative online browsing history sharing tool (Komlodi & Lutters 2008). This tool enables colleagues within the same organization to view activities of other participants, and assists novices in learning new tasks. Menczer et al. proposed a collaborative search structure that combines contextual learning and social collaboration (Menczer, Wu & Akavipat 2008). In the collaborative search, instead of querying everyone in the network, the participant communicates with only a small number of his/her neighbors. Search results using this method come from

\textsuperscript{10} http://www.digg.com
both search engines (e.g. Google, Yahoo) and other participants in the network. In addition, a system was proposed by Cohen et al. to help music consumers, especially for new ones, identify the most popular music (Cohen & Fan 2000). This web spider collects digital music download information including albums and IP addresses, and combines it with detailed information from album description to generate a personalized recommendation.

Marbach’s work (Marbach 2008) is theoretical and makes strong assumptions about uniformity of selection of resources when web-based resource access follows a strong power law. Secondly, his work proved that there exists a set which will provide full recommendation for an individual user. In contrast, we are examining the tradeoff between coverage and information-sharing in the case of recommender systems. Privacy was not a concern in Marbach’s work. He also did not address the potential for malicious participants at increased scale.

Privacy in Collaboration
Trust issues are inherent within collaborations within a social network. Kobsa (Kobsa 2007) analyzed several factors that affect opinions of online participants, and found that participants both share and care about their private information. Hawkey et al. (Hawkey & Inkpen 2005) pointed out that incidental information could be easily revealed to other participants in the same ad-hoc network. Chew (Chew 2008) found that information posted on social networking websites were available to a larger group than many people believed. These findings would be rather frustrating for participants who are active in sharing their experience online with other people. Gross and Acquisti (Gross & Acquisti 2005) showed individuals were unaware of the scope and scale of the data they share, and many desired privacy. Risks of collaboration were enumerated by Calandrino et al (Calandrino et al. 2011).

Myriad approaches to protect the privacy of participants in collaborations have been proposed. Canny designed an algorithm that calculates public aggregate data without releasing private information of participants (Canny 2002). This method is computationally efficient. Factor analysis (Canny 2002) was also used in his research on enhancing the privacy of aggregated data. Zhang et al. (Zhang, Ford & Makedon 2006) proposed a scheme with two-way communication. In that design, the server provides randomized participant perturbation in randomization, instead of using the same perturbation for all participants. To prevent the central server from collecting and releasing personal information, Shokri et al. (Shokri et al. 2009) proposed a scheme with participants’ online profile storage centralized, and offline profile storage distributed. Synchronization mechanisms were included to maintain a reasonable level of accuracy. Tsow et al. (Tsow, Viecco & Camp Accepted) proposed an architecture that stores no unencrypted information about the participants on servers, while participants submit ratings in a distributed manner. There is no perfect solution between personalization and privacy. In this paper, we quantify the tradeoffs between privacy and collaboration for the case of web search and website rating using empirical data.
Experimental Design
Recall that our research questions are twofold. “What is the optimal size of a collaborative network for evaluation?” and “What is the ideal size for discovery?” The larger research question is a fundamental query about the need for data sharing for collaboration: what is the right sharing scale for an application? We argue implicitly that designing for limited scale can decrease risk - both security and privacy risks. To answer these questions, we examined records of roughly one thousand self-similar participants over eight weeks of active use.

Data Description
The data compiled was based on the Internet browsing history of 1084 participants, who were undergraduate students living in the same dormitory of Indiana University. This is a fairly homogeneous group. We selected a dorm population based on the population itself and the network configuration. Inherently, the students in this dorm were similar in age, living environments, course loads, and current degree objectives. This was a freshman dorm populated by general studies majors. Instead of performing experiments on a defined collaborative network, we chose this homogeneous group, because our work on evaluation and discovery may not be performed on a purposely constructed collaborative network, where we could not arbitrarily manipulate group size. In addition, the construction of the experimental community was anonymous and random, so our conclusion should be easily expanded to a collaborative network with a larger population.

Browsing data were compiled from port 80, and any https traffic was not compiled. URLs were trimmed so that no session information was recorded. Participants were distinguished by a hash of their MAC addresses, and then given an identification number from 0000 to 1083 to preserve anonymity. A text file was created for each participant, where browsing activities were recorded. Every line in the file includes an http request, the visiting time, source URL, target URL, etc.

Individuals were given the option of not being included in the experiment in two ways. First, there was a VPN available for wireless connections. Obviously packet-sniffing on this from the wiring closet was not feasible. In the announcement of the experiment, the researchers’ contact information was included so that there was a two-week period before the experiment to choose to opt out. The researchers offered a seminar in the dorm for Tor installation.

A few participants’ records were removed from the experiment. In some cases, it was because their browsing histories consisted almost entirely of Windows updates. In one case, a poorly designed anonymizer generated a series of false requests putatively from other IP addresses without changing the originating MAC addresses (Duncan 2010). Overall, there are 967 valid unique participants.

Our experimental data covers a period of roughly nine weeks from March 5, 2008 to May 3, 2008. This includes a one-week school break between March 15, 2008 and March 23, 2008, when no students were in the building. We deleted this week of data. We used the first four weeks as the training set in our experiment, and the second four-week period as
the test set. In other words, we use the first four weeks to bootstrap history. We then analyze weeks five to eight. Therefore, our training and test set are obviously disjoint.

**Analysis Approaches**

Our modeling begins with the selection of a single random participant. We then increase the network size by one each round by randomly selecting a unique participant from the remaining participant pool. Based on the assumption that participants of our data collection are sufficiently self-similar to represent members of the same collaborative network, it is therefore easy to infer that our random selection also forms a community with similar interests. In each round, we examine URLs between the fifth and eighth week from the browsing history of selected participants. Each round is executed 100 times.

In the experiment, we initially assumed an experimental bound on the network size to be 50. We chose 50 for two reasons. According to the experiments performed by Ahn et al. (Ahn et al. 2007), 50 is a point where the pattern of collaborative network data becomes redundant. Our results validated this as described below, so we did not revisit that assumption. Also, according to the research done by Hill et al., the mean size of collaborative network for average person is about 125 (Hill & Dunbar 2003). Since most college students are under the age of 30, we can expect the mean size of collaborative network in our experiment to be smaller than the result mentioned in Hill’s paper.

In order to better capture the relative value of one participants’ data to another participant, we evaluate the records in three ways. The first approach is “Domain Name”, in which URLs are differentiated by the domain name only. For instance, http://host1.facebook.com/dir1 is considered to be the same as http://www.facebook.com/dir2. The second approach is “Host”, where the entire host name is considered up to the third forward slash or text termination reading left to right. The two URLs in the former example are not the same in this case. The third method is “First-Level Directory”, where we consider up to the first sub directory of an URL. The web page http://host1.facebook.com/dir1/subdir1 is the same as http://host1.facebook.com/dir1/subdir2, but different from http://host1.facebook.com/dir2/. That is, we include the field immediately to the right of the third forward slash, reading left to right.

For the evaluation of sites, the sites are identified as either “Known” or “Unknown”. This method of evaluating known websites uses a time “threshold”, which is based on previous research showing that phishing websites are usually taken down after a short period of time, malware-distributing sites are identified in the same window, and thus both can be distinguished from a non-malicious site (Moore & Clayton 2007). Therefore, a site is labeled “Unknown” if at the time point of a browsing record, there is no former record (from the beginning of the first week to the time stamp of the record in the second half of the eight-week period). The records search includes not only the single participant to whom the current URL belongs, but also the entire collaborative group. For example, suppose participants A and B are currently in our participant pool, and the threshold is seven days. Then if participant A visits the URL http://www.website1.com for the first
time, this URL should be judged as “Known” as long as user B saw that site seven or more days ago.

Please note that the browsing histories of the selected participants in the first four weeks are treated only as history, so these are not included in the final coverage rate calculation in our results. For example, if someone visited the URL http://www.website1.com in the fifth week, which is in the test set, and the first visiting time for that URL was in the first week, which is in the training set; the URL would be considered as “Known” in the fifth week as the threshold rule is met. However, if the site was not visited on or after the fifth week, it is not included as having been rated for the purposes of communal evaluation.

In the experimental analysis, we use seven days as the threshold for judging a URL to be known or not. We base this on previous empirical analyses of malicious websites. Although some research suggested that it took up to two weeks to take down a phishing site, APWG has reported that the average uptime for the sites has dropped to as low as three days (APWG 2010) due to both defenders shut down sites more quickly and attackers create multiple short-lifetime phishing sites. Moore et al. (Moore & Clayton 2007) showed from their empirical analysis that the average lifetime for a normal phishing site is about 56 hours. In addition, considering patterns the browsing history possesses, seven days should be an appropriate length for the threshold. For example, suppose a biology class meets every Monday, and many students view course webpage and the virtual laboratory only on that day. Similarly, on-campus jobs post every Friday, so we can expect a large number of visits to the online message board to occur only when it updates each week.

Another important issue in our analysis is the strategy for adding participants. In our first analysis, for each round, one new participant is added into the previously constructed collaborative network. Suppose participant 0001 is in the pool during the first round, then we add one new participant, say 0139 into the participant pool in the second round with participant 0001 remaining. In our second analysis, we re-initialized the collaborative network instead of adding incremental participants. In other words, the group is entirely reconstituted for every n in [1,50]. Considering the same case in the second round, we remove the first participant 0001, and pick two new participants. The participants for the second round might be 0139 and 0002. That is, we sample with and without replacement. Results for these two methods are similar in our experiments. When a new participant is added into the participant pool, his or her browsing history is combined with rest of the participants, and each participant’s history is treated equally. For example, suppose participant 0002 is added into a pool that already contains participant 0001, then in the next round, we need to consider all records for 0001 and 0002 in the first four weeks as training set, and all records of the two participants that occur in the second half are in the test set.

**Data Analysis**

We examine this data at three different depths as described above (i.e. domain name, host and first-level directory). In order to resolve the apparent conflict between assumptions about participant search space versus participant advice networks, we analyzed according
to both clicks and sites. Each record in the participant file corresponds to a click. Each site may be visited thousands of times. The coverage rate for known sites is not equivalent to rate for known clicks by definition. For each approach, we calculated the arithmetic mean for the 100 runs of the experiments per network size, and the maximum and minimum rates were also recorded in separate graphs. We used this Monte Carlo approach to prevent sampling idiosyncrasies from affecting the overall results. We examined both to provide a view of participant history (i.e. sites) and participant experience (i.e. clicks).

Figure 1 shows the average coverage rate when we consider the first network construction strategy (i.e. adding one participant per round). For clicks, the results illustrate that the first five people added to a network vastly increase the value of the feedback for communal evaluation. That is, if a member of the collaborative network has not visited a site, then they could not have evaluated it. Indeed, visiting itself is an evaluation. After ten participants are added, more than 95% of the domains have already been covered, while host and first-level directory also exhibit coverage values of nearly 90%. At that point, however, the average value of each additional participant brings very little increase. After 40 additional participants, nearly 99% of clicks will go to known sites. Recall that a known site has been previously visited, which is itself an implicit rating (e.g. attention span was expanded) and the site has existed for at least one week. The minimum coverage shown in Figure 2 has a trend similar to the mean values, with only 5% less than the mean value. Figure 3 shows the maximum coverage rate for clicks among the 100 rounds, and the drop of clicks at the lowest level is a result of sampling with replacement. Notice that the intersection would grow more slowly with a more heterogeneous population, so our findings may hold more strongly but with a network size greater than five.

The value of an additional participant grows rapidly only until a small number (in this case, 5). After the rapid growth in coverage, the value of inviting more participants in the browsing history sharing decreases. Recall that adding more participants into our network would result in a higher risk that private information could be released. By measuring the number of clicks that can be informed by previous visitors in the collaborative network, we can determine how much coverage in terms of total online experience can be provided by evaluation as a function of collaborative network size. Participants' experience with a reputation system will arguably be a function of clicks. Three percent of clicks resulting in no evaluations is a very different experience from 3% of sites having no evaluation.

Discovery is valuable for the entire space covered by the search, or union of participants' histories browsing history. Figure 4 shows the additional evaluation space and expansion of the search space created by adding a second user, while Figure 5 illustrates the expansion in search and evaluation spaces when a third user is added. By measuring the number of sites that are distinct we can determine the relative size of the expansion of the collaborative search space. Recall that, in contrast, evaluation requires an intersection of experience or sites that are shared by participants. In contrast, search space is the disjoint elements in each additional participant’s history.
We use “clicks” to observe the relation between network size and coverage rate in communal evaluation web services. Recall that ratings for different websites are generated from historical browsing records of all participants in the group in an experimental round. Therefore, the complete browsing history (includes participant ID, visiting time, and URL) is required for an accurate result. In contrast, collaborative search web services focus mainly on discovering the web, so it would be sufficient to include only the URL itself in calculations.

For sites, the results illustrate that there is a low and roughly constant expansion of web coverage as the number of participants in a collaborative network increases. We do not address malicious network participants in our analysis. The average marginal value for the addition of an n\textsuperscript{th} website remains significant and positive as the collaborative network increases as illustrated in Figure 6. Note that this union of identified resources would grow more quickly in a heterogeneous population, and an increase in network size would have a larger but roughly constant increase (on average).

![Figure 1 Average Coverage Rate for Clicks.](image-url)
Figure 2 Minimum Coverage Rate for Clicks.

Figure 3 Maximum Coverage Rate for Clicks.
Figure 4 Expansion of Evaluation and Search Spaces (Two users).

Figure 5 Expansion of Evaluation and Search Spaces (Three users).

Figure 6 Average Coverage Rate for Sites.
Conclusion
Collaborative networks are often designed with the assumption that an increase in the number of participants is always good. However, when considering privacy and security, it is clear that there are risks to ever-greater participation. In this work, we have clearly illustrated that the benefits of the tradeoff between participation and privacy are decreasing in one application and at best constant in another. Collaborative search (i.e. discovery) assumes that the members in a collaborative network have large sets of non-overlapping sites so that an increase in network size corresponds to an increase in web coverage. In contrast, communal evaluation (i.e. experience) assumes that experiences are self-similar, so that a site one person visits is likely to have been viewed and thus potentially evaluated by others. Our research has shown that this is not, in fact, a discrepancy. In terms of search, when seeking a site, the largest possible collaborative network will result in the largest possible coverage. There is a roughly constant marginal rate of increase between three and fifty additional participants in our finding. If we assume the costs (including the risk of including a malicious or dangerously incompetent participant) are constant, then search shows a roughly constant benefit. If these costs increase, then the benefit/cost ratio is also decreasing in the case of search. In this paper, we quantify the tradeoffs between privacy and collaboration for the cases of web search and website rating using empirical data. The most straightforward way to manage trust in a collaborative network may be to limit the number of participants based on the function of the data sharing.

Evaluations are sought to enhance one’s experience while search enables discovery. The difference between sites and clicks illustrates that collaborative networks for discovery and evaluation are ideally discrete but complementary. Evaluations can be available for 95% of a participant’s click experience with 10 participants and 98% with 50. There are decreasing marginal returns for evaluation, without considering the potential for untrustworthy or malicious actors with a larger network. Evaluations are needed for clicks that are shared, i.e. the intersection of the experience. Search is needed, by definition, for sites that are known and unknown, i.e. the disjoint component of the participants’ experiences. After three participants, the union grows at monotonically decreasing but positive rate. Increasing the participant set continues to increase the union. Thus, the set of participants optimal for evaluation may be a very small subset of that for discovery. Evaluation networks can be smaller, providing fewer evaluations from socially trusted evaluators or peers. Evaluation also requires offering more detailed information (e.g. opinions or patterns) than discovery (which requires only knowledge of a resource’s existence).

To summarize our findings: when it comes to discovering new websites, an expansive collaborative network is ideal; however, when considering what you trust, a network over ten offers decreasing marginal returns. Treating these problems as discrete may simplify privacy-enhancing design for evaluation and discovery as well as inherently addressing issues of trust.
Bibliography


Marbach, P 2008, 'A lower-bound on the number of rankings required in recommender systems using collaborative filtering', *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*.


Metcalfe, R 1995, 'Metcalfe's law: A network becomes more valuable as it reaches more users', *Infoworld*, no. 17.


Park, S-T & Pennock, DM 2007, 'Applying collaborative filtering techniques to movie search for better ranking and browsing', *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*.


