Collaborative Search and User Privacy: How Can They Be Reconciled?

Thorben Burghardt[1], Erik Buchmann[1], Klemens Böhm[1], and Chris Clifton[2]

[1] Universität Karlsruhe (TH), IN-F, 76131 Karlsruhe, Germany
{burgthor,buchmann,boehm}@ipd.uka.de
[2] Dept. of Computer Science/CERIAS, Purdue University, West Lafayette,
IN 47907-2107, USA
clifton@cs.purdue.edu

Abstract. Collaborative search engines (CSE) let users pool their resources and share their experiences when seeking information on the web. However, when shared, search terms and links clicked reveal user interests, habits, social relations and intentions. In other words, CSE put privacy of users at risk. This seriously limits the proliferation and acceptance of CSE. To address the problem, we have carried out a qualitative study that identifies the privacy concerns of CSE users. In particular, our study reveals the range and type of concerns when sharing query terms and search results with different social groups, e.g., family members or colleagues. To control the information shared, the participants of our study have called for anonymity and reciprocity in combination with time- and/or context-dependent conditions. To facilitate the specification of privacy preferences, we define a general policy structure to express privacy needs in the context of CSE. We also give an approach to address the reciprocity condition identified in the study, and we discuss options to anonymize sharing of query terms.

Keywords: Collaborative Search, Privacy, Policy.

1 Introduction

Collaborative Search Engines (CSE) enhance web search by sharing query terms, search results and links clicked among users. Examples include I-SPY [1], MUSE and MUST [2], SearchTogether [3] and Fireball LiveSearch [4]. CSE let knowledge workers synchronize efforts, provide guidance for inexperienced searchers, and offer Web-2.0-style information on Internet activities of friends/colleagues. So far, collaborative search has been done by hand [5], e.g., by sending emails with search results. CSE are more efficient in this respect. However, as queries and the results clicked can reveal the habits, interests, social relationships and intentions of the searcher [6], CSE are problematic from a privacy perspective.

Current CSE either state that any search information will be visible to others, and/or leave it to the user to manually invite individuals to benefit from a particular piece of information [3]. CSE require that information is shared automatically, or that people can subscribe to information generated by others, in

E. Bertino and J.B.D. Joshi (Eds.): CollaborateCom 2008, LNICST 10, pp. 85–99, 2009.

the spirit of friendfeed.com. Acceptance of such an environment will depend on the appropriateness of privacy mechanisms. It is challenging to determine what constitutes a suitable privacy mechanism. CSE are about supporting other individuals; privacy laws that regulate how organizations can collect and use private information do not really fit. Web privacy mechanisms such as P3P [7] focus on the user-provider relationship, while CSE support sharing between individuals; the privacy needs are likely to be different. Also, privacy is a highly emotional issue; previous studies have shown that humans do not necessarily reveal their true privacy needs in laboratory experiments and/or questionnaires [8].

To tackle these challenges, we have conducted a user study with 27 computer science graduate students. Therefore, we constructed a CSE with mock-up privacy mechanisms to observe the true privacy needs of the users. Since CSE are a new technology, we familiarized the participants with CSE concepts and usage. The mock-up privacy mechanisms let the participants specify their privacy preferences for collaborative search in natural language. Asking for policies in natural language rules out any technical limitations, e.g., limited expressiveness of a specific formal language. Questions we wanted to answer include: (i) What are the **parameters** of polices, e.g., “at work”, “not before 8 pm”, used to control disclosure of search information? (ii) Which **groups**, e.g., friends, family or colleagues, do CSE users address in their policy definitions? (iii) Is there a common **structure** of the policies so that they can easily be transformed to a machine readable representation? (iv) Do the users express different privacy needs for the query terms and the links followed from the query result?

Our evaluation yields interesting insights (described briefly in [9]): Individuals are concerned most about what friends, colleagues and family might learn from their queries. This is particularly interesting given that SearchTogether [3] found that these are the people individuals want to collaboratively search with. We also found that participants express their privacy needs with policies of simple structure, but refer to different kinds of constraints. Our participants also defined reciprocal conditions, i.e., they take characteristics of other users into account, like having similar search interests, or distinguish between users registered for a long time and new ones. This is noteworthy, as we are not aware of privacy preference specification languages that incorporate this concept. Furthermore, most users do not distinguish between sharing query terms and sharing links.

Finally, we outline how a CSE supporting privacy could be constructed. Some of the expressed policies would be difficult to guarantee if the CSE manages privacy. This is because the authority that would control the CSE may be the very authority to whom some policies prevent disclosure. We describe techniques for decentralized search, collaboration, and policy management that could be used to overcome this limitation.

Paper structure: We review related work in Section 2, and we describe the methodology and key design decisions of this study in Section 3. Section 4 features the study results together with a policy specification language that can represent the user requirements. We present approaches to enforce these policies in Section 5; Section 6 concludes.

2 Related Work

While people are willing to share information [10], they distinguish between individuals, e.g., friends, relatives, or colleagues [3]. General studies have shown that approximately 90% of participants are concerned about privacy [11]. Privacy studies, e.g., [12], do not focus on the unique traits of CSE. Technology-independent studies [10,11] reveal the general willingness to share personal information if privacy mechanisms are available, but do not give us any hints on how to design such mechanisms for CSE.

User studies [13,5,14] have shown that individuals already collaborate for complex search tasks, e.g., holiday planning or homework; [5] shows that more than 85% of their "relatively sophisticated web searchers" share results of a web search, with over 25% cooperating on a weekly basis and over 75% monthly. They share using approaches such as email or instant messaging; the "push" nature of such sharing allows users to implicitly enforce their privacy policies; these studies provide little insight into user privacy requirements.

While some CSE exist, e.g., I-SPY [1], SearchTogether [3], MUSE [2], or Fireball LiveSearch, we are not aware of any research on the privacy issue. For instance, SearchTogether shares query histories, personal comments on pages and information on pages visited automatically, without any restriction regarding privacy. The design goals of SearchTogether are *Awareness*, *Division of Labor* and *Persistence*. Awareness of the search processes and search results of others means that people can learn from experienced searchers and work together on a search project without explicitly asking for information on pages visited or search terms tried; this prevents the automatic/implicit self-enforcement of privacy preference in push-based collaboration. SearchTogether makes search sessions persistent, including the query history, links clicked and comments provided. Evaluation shows SearchTogether to be more efficient than conventional methods. A study of MUSE, which has a similar structure, has focused on communication in the context of CSE [2] and has shown that users frequently wish to communicate. SearchTogether, MUSE and I-SPY let the user manually select which information should be shared with whom, based on search sessions or groups. While this allows users to enforce their privacy policies, it also requires them to manually ensure that they do so: within a session, any user (even one joining later) can see the whole search history.

People also share information with open communities. With the Fireball LiveSearch, the search engine displays all query terms it is currently processing. There are around 250,000 visits of Live Search per month. Fireball satisfies the curiosity of others, it allows users to learn, and it provides suggestions for future queries. The information displayed is not restricted to certain topics or users.

Few existing single-user search engines feature privacy mechanisms. The AskEraser of Ask.com deletes all information on past search activities and turns off all personalization features. This involves a tradeoff: Simply deactivating logging increases the privacy of the users, but can reduce effectiveness of the search [15,16,17,18].

3 Study Design and Environment

We now describe the key design decisions behind our study, and the environment developed to support these decisions.

3.1 Study Design Decisions

Skilled Participants. Understanding the information flow and ways to build privacy profiles for queries and links shared requires a thorough knowledge of information systems. Our study participants are graduate students in computer science with a focus on information systems, making them skilled Internet users and knowledge workers, a target group of CSE [3].

Training on CSE. Currently, only research prototypes and early implementations of CSE exist; we did not expect participants to possess in-depth knowledge of CSE. As competence is required to obtain meaningful study results [19], we implemented a CSE prototype ourselves and ensured that participants have familiarity with its use and technical details. To avoid influencing the results of the study, we did no training on privacy threats or privacy-enhancing technologies.

Observing the Behavior. Since users adapt their behavior to the technology available [19], it is problematic to obtain real privacy needs by means of abstract questionnaires or synthetic experiments. We implemented mock-up privacy mechanisms that allow us to observe the privacy needs of participants working with an operational CSE.

Plaintext Policies. A final design decision was to let the participants specify policies in natural language instead of using a machine-readable policy language. First, this does not restrict the expressiveness of policies. Second, it is intuitive; if the participants had to learn and use a formal language unfamiliar to them, the added effort could limit the number of policies obtained. Third, use of an open-ended natural language interface ensures that the mechanism used does not influence the policies specified, supporting our goal of learning all criteria users find important to privacy in CSE.

3.2 Study Environment and Methodology

We now briefly outline the methodology used for our study, including the components of the CSE and the training approach used to satisfy the design decisions of the previous section. Today's CSE consist of three main components [3,2]: a search engine where users can enter their information needs, an integrated mechanism that allows users to exchange query terms as well as links clicked and to make them persistent for future search sessions and other users, and a way to communicate between the collaborating parties. In the first three of five three-week phases, participants implemented a portion of each component (as projects in a database course) to ensure familiarity with the technology.

Introduction Phase (P1). To generate a basic understanding how CSE work, we prepared a presentation introducing the functionality and let the participants search for common topics individually. We built our CSE on top of Google, because our participants were familiar with the use, layout, and quality of the search result. Furthermore, [1] showed that the effectiveness of the collaboration strongly depends on the reference search engine; they achieved best results using Google. The queries and query metadata were stored in a relational database; during Phase 1 the participants wrote small programs to access the search data, such as SQL queries to find searches similar to their own.

Query Phase (P2). We then activated collaborative aspects of the CSE, specifically a window that displays similar query terms (calculated using an Unweighted Vector Space retrieval model [20]) and clicked links without the name of the issuer (as with I-SPY [1]). The participants used these to find information to solve the training tasks. We also enabled logging of clicks on similar queries and links. As working with the database was part of the training task, the participants knew in detail what information was stored and processed by our CSE.

User Awareness Phase (P3). This phase incorporated user identity into the interface (previously seen by participants through access to the database as part of their training tasks.) The participants developed a Skype plugin to subscribe to query terms of other participants, giving subscribers a notification containing the query term, links clicked, name of the issuer, and the time of the query. This interface enabled messenger-based collaboration when searching. The Skype feature "asynchronous communication" was leveraged to cache messages when others are offline; the history function gave persistence of the communication.

Policy Definition Phase (P4). We then asked participants to enter their policies using a plaintext (natural language) policy editor. When using the CSE, users selected a (self-defined) policy (e.g., "being at work") from a list; if a policy is selected, the current query term and links clicked are available only for users that match (as implemented by study administrators based on the natural language policy.) The policy stays active until the user switches to another one.

While we did suggest that policies should encompass in *which context* they would or would not share *which query terms and clicked links* with *whom*, the policies were specified in natural language to give users the freedom to realize their own notions and ideas when specifying their privacy needs. As policies can be sensitive [21,22], we kept all policy definitions private. To avoid overtaxing the participants and to allow for a structured evaluation, this phase had three steps: 1) policies to protect query terms, 2) policies to protect links clicked, and 3) three weeks of CSE use to give the opportunity to refine policies.

Survey Phase (P5). We closed with a survey 1) asking the participants control questions, to guarantee the representativeness of our CSE, and 2) to obtain information about their general privacy attitude. This information was used in interpretation of our results. To motivate participation in the final survey we drew two Amazon vouchers among the participants; participation in phases P1–P3 was ensured as the tasks were graded course requirements. We did not give grades or inducement on defining policies (P4) to avoid influencing the results. Supplementary information on the CSE and study methodology is available on an accompanying web page [23].

3.3 Study Representativeness

Before describing the outcome, we outline the background of participants and why we feel the results are representative of typical CSE users. The 27 graduate students in the study were enrolled in a practical course in database systems at the University of Karlsruhe. All participants had a fundamental understanding of information systems, complex search tasks, and are interested in data analysis and KDD; but as the goal of the course was knowledge of database systems, we did not expect them to be particularly biased toward or against CSE. The students represent a range of cultural backgrounds, covering seven nationalities (ten German, six Hungarian, four Bulgarian, three Chinese, two Ukrainian, and one each from Belorussia and Romania.) We had 7 female and 20 male participants, with age ranging from 20 to 34 years (avg 24).

To gauge the privacy attitude of our participants, we asked their usual practices regarding privacy policies, registering at search engines, and querying personal information on the Internet. 66% said they have read the privacy policy of at least one web site, 27% stated they read them frequently (but not always) when registering. We found it interesting that 81% said they had never read a search engine privacy policy while 78% have registered with a search engine for further services like email or a messenger account. 59% expected that their search engine can link their identity to each of their queries. The results indicate that participants had no extreme privacy attitudes, and while not naïve, are probably not fully aware of privacy threats in the context of search engines.

To evaluate if our prototypical CSE is representative and if the participants represent individuals likely to use CSE, in Phase 5 we asked the participants how they perceived its components. On a 5 point Likert scale over 70% found our CSE and the links proposed medium (10), useful (9) or very useful (1). Asked if they would use the CSE for a learning group, three stated no, four rather not, but 67% stated rather yes (15) or absolutely yes (3). We see this as a confirmation that the CSE is useful and that the study has not been biased by technical limitations. 59% (16) of the participants have investigated query terms of others by browsing the search history of specific users. Reasons for those who did not are “no interest in searches by others” (4), “finding Google recommendations sufficient” (3), “finding searches by others not good enough”, or “no need for this functionality” (1). Four participants stated that the effort necessary to select a policy before a search was very easy (grade 1), 6 gave a 2, 9 chose the middle, and 4 each chose grade 4 and 5. This gives us confidence that participants felt the privacy mechanism realistic.

We conclude that participants deemed our CSE useful. Many would continue to use it, and are willing to share query information within the constraints of privacy preferences. Thus, we can expect realistic results from our study.

4 Study Evaluation

We now describe the results of our study, i.e., we analyze the policies provided by our participants. We obtained 247 policy definitions from the 27 users. 142 policies consider sharing of query terms, and 105 address the links clicked. We first analyze the policies assigned to the query terms. We investigate (1) how these policies are structured, (2) the contexts they refer to, (3) which social groups are mentioned, and (4) the form of the policy predicates. We then investigate how the policies for the links clicked differ.

4.1 Policy Structure

We found that the plaintext policies of the participants can all be expressed using the following general structure:

[ALWAYS | IF <*conditions*>] [DO NOT] DISCLOSE <*objects*> [TO <*groups*>]

Conditions, objects and groups were composed of one or more terms connected with AND or OR, e.g., "If I am at work OR time between 7 am and 6 pm DO NOT DISCLOSE query terms to my friends AND family". Some policies followed this structure literally; the rest could be transformed to fit the structure.

While some policies allow or prohibit information disclosure without conditions ("ALWAYS DO..."), most refer to one or more of the following:

1. *context* (e.g., "while I am at work")
2. *content* ("the query contains adult material")
3. *time* ("between 7 am and 6 pm")
4. *reciprocity* ("if the other user has similar query terms")
5. *query-result dependency* ("show the query term if the clicked link refers to a newspaper site")

The difference between context and content condition is that the former relates to the user, the latter to the wording of the query. In the policies formulated by the participants, the *object* can either be a query term, a clicked link, or both. The *group* specifies individuals that may/may not access a certain object.

The conditions, addressed groups, and objects are orthogonal to each other, i.e., we did not see one group mentioned only in combination with a specific context, etc. Therefore, we now evaluate these aspects separately.

4.2 General Policies

The simplest variant of policies allows or forbids the disclosure of the query term without specifying conditions or persons, similar to "ALWAYS DO NOT DISCLOSE anything TO anybody". 12 of our 27 participants created a policy that always prohibits the disclosure of the query term, 5 defined a policy that discloses the query term to everyone. This is similar to studies on other technology; 90% of the participants in [11] are moderately or very concerned about privacy.

4.3 Conditions

For policies with a condition, 39% (56) refer to a context, 11% (16) to the query term (content condition), 7 to characteristics of the person who wants access (reciprocal condition), and 3 are time-dependent (time condition). In general, policies were simple. The vast majority had only one condition. A few combined (at most two) different conditions, e.g., "If I am at work, and the time is between 7am and 6pm". Policies did not differentiate between sharing the query term and the metadata of the query such as query issuer or the time the query was issued, although one participant did express a desire for anonymous sharing.

Contexts. The context describes the current situation of the query issuer, e.g., at work, planning holidays, etc. 39% (56) of the 142 query term policies can be assigned to a concrete context (see Table 1; the accompanying web page [23] contains examples of each group.) The remaining 86 policies apply to all contexts.

Table 1. User Contexts

Context	Frequency
Being at work	21
Private surfing at work	10
Searching for adult material	5
Searching provider related content (e.g., "youtube videos")	5
Online shopping	4
Searching for disease	3
Searching for jobs	2
Planning holidays	2
Searching for dating sites	1
Searching for person names	1
Searching for sports issues	1
Money management	1

Participants defined policies for many contexts not explicitly addressed by law, e.g., by the EU directives [24]. Laws typically specify contexts relevant to information sharing between an individual and an organization, such as medical issues or employer-employee relationships. Our participants also specified more personal contexts such as holiday planning or searching for persons of interest.

Content Conditions. Content conditions refer to the query term. However, as the content of the query and the context that motivated the user to issue that query are closely related, some policy definitions overlap. 11% (16) of all policies defined at least one policy comparable to "*If I issue a query containing <some keywords>, (do not) show the query to <some persons>*". 5 policies refer to a provider name, e.g., youtube or newspapers, 3 refer to person names, 3 to technical issues and 2 to sex and porn. See [23] for the full list.

A rather surprising content condition relates to the structure of the query. Participants stated 5 policies where query terms are (not) published if they consist of less (more) than a certain number of words. We conclude that the participants intend to restrict the level of detail, i.e., they assume that longer search terms carry more or more sensitive information than shorter terms.

Time Conditions. Three policies from different users define conditions based on time and date. For example, one policy forbids disclosing search information during working hours (this was to prevent competitors benefiting from the person's work.) Another participant allowed sharing query terms with friends only between 6pm and 8pm. One policy used date; it forbid sharing search terms with friends before Christmas while shopping for presents.

We were surprised that no participant generated a policy that takes the sequential order of the queries into account or requires a certain delay between the time the query was issued and the time the query is shown to another person. Instead, the participants preferred to either share the query terms at once or prohibit access completely. As with ALWAYS (DO NOT)-Policies, this is another hint that users want to keep policies simple.

Reciprocal Conditions. Reciprocal conditions depend on characteristics of other persons when deciding if a certain piece of information should be disclosed. Three users defined a total of 7 reciprocal policies: (i) Five share query terms if other users have previously issued similar queries, e.g., by requiring a number of identical words in the query term, (ii) One allows sharing the query term with users who registered to the CSE before the issuer of the query, and (iii) One participant requires that a query can be shown to someone else as long as this other person does not learn the issuer's identity from the query term.

The results indicate that some users are willing to share information with like-minded people only, e.g., if they suffer from the same disease. Further, one participant explicitly called for anonymization, i.e., sharing query terms and links only if identity is not revealed. Reciprocity conditions indicate that other technologies cannot readily be transferred to CSE. For example, no formal privacy-preference language we are aware of considers reciprocity conditions.

4.4 Groups

A group defines which individuals are allowed (or prohibited) to see a query term. We were interested which social relationships and classes of social groups (like *friends* or *family*) participants address in their policies. One insight is that social groups can be divided into (i) groups containing only individuals which are personally known or (ii) groups with unspecified members. For example, *family members* are known personally, while *children* stands for an unspecified group. Our participants defined 106 policies referring to various social groups, cf. Table 2. 60% (82) relate to Class (i), 40% (42) to Class (ii). 83% (88) policies only address one group per policy, 17% (18) policies address multiple groups.

The most common group in Class (ii) was "colleagues". This was not surprising given policies that address topics like "job search" that one would expect need protection from colleagues that are not close friends. Regarding Class (i),

Table 2. Social Groups

Group	Class	Frq	Acc.	Prohib.
Friends	known	35	8	27
Family	known	19	8	11
Acquaintance	known	12	7	5
Boy/Girlfriend	known	4	1	3
Supervisors	known	4	3	1
Doctors	known	3	0	3
Teacher	known	3	3	0
Parents	known	1	0	1
Landlord	known	1	1	0
Colleagues	unspec	26	6	20
(Fellow) Students	unspec	6	2	4
Children	unspec	3	3	0
Male/Female	unspec	3	0	3
Official Persons	unspec	3	3	0
Fellow Citizen	unspec	1	1	0

participants addressed friends (33%) and family (18%) most often. 20 of 27 (74%) participants have specified a policy that allows sharing queries with friends.

Further, we explore if these groups are used to restrict or increase the set of individuals allowed to see a query. We differentiate between policies used to grant access to information, e.g., "*show my query to my colleagues and friends*", and groups used for the reverse, e.g., "*do not let my boss see my query*". In our study, participants used both variants frequently and even combined them, e.g., "*give access to my friends but not to my family*". 64% (68) of our 106 group-based policies grant access, and 36% (38) use groups to prohibit access to query information (Table 2). Our study indicates that individuals are concerned about what friends, colleagues or family members might learn from their query terms. This is interesting: [3] shows that it is exactly these individuals people want most to search with collaboratively.

4.5 Link Policies

With our setup, the object specified in the policy can be a query term or the link clicked. The 142 policies analyzed so far address the query terms; we now investigate the 105 policies that refer to the links clicked. 81 policies on links are copies of policies referring to query terms. 7 new policies have been defined. 17 policies originate from policies on queries, but have been extended. Three participants have not specified any policy on links clicked.

Those 23% (24) that differ are either general *ALWAYS DO NOT*-policies, policies that have been further restricted, e.g., by specifying additional *groups* that may (not) see the links, or define *content conditions* on the URLs or descriptions of the websites displayed as part of the query result. Interesting are 6 policies with a *time condition*: 3 require that information on links clicked be disclosed to others for some days only. These policies allow participation in the CSE, but rule out the derivation of long term profiles. Three policies allow sharing the last n links clicked only. Four policies define so-called *query-result dependencies*. Such dependencies are a new class of conditions: They share query terms depending on the links clicked, e.g., “Do not share query terms except when a link clicked leads to a newspaper site”.

4.6 Discussion

We found that policies fit a simple structure consisting of one or more conditions, objects and groups of persons. Policies do not contain conditions from more than two different classes. Participants have created very specific policies, e.g., in order to prevent family members from learning about presents before Christmas, as well as abstract policies like “prohibit any access”. Since we did not discuss possible privacy threats or existing privacy-enhancing technologies, the set of policies is probably incomplete. However, the policies reveal the spectrum of requirements that users deem useful, i.e., a privacy mechanism should at least comprise. This includes lists of keywords, although providing and maintaining a comprehensive list of sensible words is a daunting task. WordNet [25] or other thesauri could help to address word concepts instead of individual words. Policies also refer to social groups that roughly correspond to the social relationships of the query issuer. One approach to simplify group definition could be extracting relationships from social network sites or messenger services [26]. Identifying unspecified members of groups, e.g., colleagues or children, without raising new privacy threats by additionally revealing data like age or employer, will be much more difficult, although issues of such group membership are a problem for identity management in general. Further, policies address time and content constraints, reciprocal constraints that take into account characteristics of other users, and distinguish between query terms and links clicked.

5 Enforcing CSE Policies

A CSE provider learns much about each user. This is particularly critical if the provider not only supports collaborative search, but also manages user privacy. Suppose a CSE in a corporate setting to support collaborative search among an engineering team. Asking the corporation CSE to enforce the policy “do not let my boss see my query” is problematic.

In this section, we assume a distributed, anonymized system architecture, and we propose mechanisms to manage policies at the client. Each of the CSE components identified in Section 3.2 can be realized by using existing tools and techniques, e.g., anonymous instant messaging such as TorChat, anonymous information sharing such as FreeNet, and anonymous web search tools such as PWS [27]. This allows separating collaboration from the search engine, thus allowing use of public search engines while supporting collaboration on a local server only when user policy allows, or even through a peer-to-peer network.

Most policies can be enforced locally at the user’s machine: The policy context, content, time, and query-result dependency either allow the query to be shared or not. Two types of policies cannot: reciprocity and anonymity. We now describe how these policies could be enforced through collaboration between parties, enabling privacy preference handling without the help of third parties.

5.1 Reciprocity

Reciprocity means that a user wants to share only if the recipient has certain characteristics, e.g., a compatible policy or similar interests. Complicating this is the fact that the conditions themselves may be sensitive. For example, a law enforcement officer may only be willing to share searches about a suspect with other officers who already know of the suspect; revealing the suspect to determine if other officers are searching for that suspect is nearly as compromising as revealing the search. In the following, we focus on typical reciprocity policies that require both parties to be willing to share similar content, while revealing query terms to determine if other parties are willing to share inherently violates that policy. Extension of the approaches presented to conditions such as "share only with people who have issued similar queries in the past" is straightforward. We sketch how this could be accomplished for reciprocity conditions requiring equality tests on context, content, time, and query-result dependency; for conditions that are too complex to efficiently map into a set of equality tests, trust negotiation approaches can be used (e.g., [28]).

Using commutative encryption, reciprocity conditions involving exact match of conditions can be tested in a peer-to-peer fashion that ensures nothing is learned except the conditions that match. The basic idea behind commutative encryption is that $E_a(E_b(m)) = E_b(E_a(m))$. If m is a tuple consisting of context, content, time, and query-result dependency, each party encrypts the tuple m with its own key, then passes the encrypted tuple to the other party. Each party then encrypts the encrypted tuple with its own key; the parties can now share the (doubly) encrypted tuple. If the doubly-encrypted tuples are the same, then the conditions match (see [29] for more details and proof that nothing is revealed.)

In practice, the users have to compare sets of policies. For example, "share queries using these terms only if the recipient agrees to sharing similar queries." means that multiple combinations of the terms specified must be compared. Our approach can be used to find a set of matching policies, then both parties can decrypt the matching policies to expose them (note that both parties must agree to and participate in the decryption.) Alternatively, the test can be performed on a per-collaboration basis: One party encrypts all relevant policies and the query terms relevant to that policy, the other encrypts the particular policies and terms that apply to the particular collaboration. If there is a match, then the parties have reciprocity.

5.2 Anonymous Collaborative Search

One user explicitly called for anonymous sharing. This is challenging, as the query terms and links may inherently be identifying [30]. We propose a model based on k-anonymity [31,32]: users agree to share query terms if at least k users have issued a query with the same terms. This can be checked using the same commutative encryption ideas above, see [29] for a more secure approach.

Note that everyone needed to form such a group of k users must have a policy of only sharing those terms/links anonymously. Otherwise, a query could be used to form such an anonymous group, then the user could reveal they had issued it; then only $k - 1$ users would remain anonymous, violating their anonymity constraint. However, it is likely that the types of searches where some users desire anonymity will be ones where many will (e.g., non-work-related searches performed at work, politically or legally sensitive topics), so this is likely to be a reasonable constraint in practice.

If any term in a query is covered by an anonymous sharing policy, all terms that are shared must meet the k constraint, not just the content condition in the policy. This is because terms not covered by the policy may be the ones that are inherently identifying. For example, one could say that a query "computing jobs available" could be revealed only anonymously; then issue a query "computing jobs available near Fasanenplatz". Even if k users issued "computing jobs available" queries, it would not be safe to disclose the query including address unless at least k individuals had issued a query with the same address.

It would be possible to share only the subset of query terms / links clicked that meet the k constraint. As with reciprocity, techniques from distributed privacy-preserving data mining (in particular, [29]) can be used to anonymously determine when sharing is possible in a peer-to-peer fashion. This is similar to the commutative encryption approach above, except that a final cryptographic protocol is used to disclose only if the number of users issuing each term meets the k threshold. Further study is needed to determine:

- How large a community of collaborators is needed to generate a pool of identical query terms sufficient to meet the k constraint? This study had 23 participants who generated at least one query, with an average of 61 queries per user. The queries averaged just under three terms. We had five 3-term and 26 2-term subsets that were 3-anonymous, and 11 4-term, 49 3-term, and 138 2-term subsets that were 2-anonymous (for comparison, there were 277 distinct 3-term queries and 391 distinct 2-term queries.) The users did have common tasks, so overlap is to be expected; this would be likely in envisioned collaborative search environments such as within a company. While a larger user base is needed to obtain a reasonably high percentage of anonymous queries, we can see that k-anonymous collaborative search is plausible.
- Is sharing a subset of a query as effective in supporting user search as sharing an entire query? (Eleven of the 49 3-term 2-anonymous queries exactly matched real 3-term queries, the rest were subsets of larger queries.)
- Does anonymous sharing provide value, or is knowledge of who has performed a search necessary to give credibility to the process?

6 Conclusions

Collaborative search engines (CSE) are an important new trend in Internet search. Information shared by CSE can put privacy of users at risk. To gain insight into this important issue (the privacy needs of CSE users), we implemented a CSE and used it during a one-semester course to give 27 study participants a thorough understanding of CSE technology. This let us observe the real privacy needs of users with an operational system.

While a few individuals define "don't care" policies, most define policies for various contexts addressing different social groups. The groups friends, colleagues and family most frequently addressed in policies are the ones people want most to search with collaboratively. This underlines the importance of privacy mechanisms for CSE. Further, individuals make use of different conditions in their policies but tend to keep the policies simple. Some users call for reciprocal conditions that depend on characteristics of others. This is noteworthy, as we are aware of no privacy approaches that consider this issue. Fortunately, these needs can be addressed: we have outlined a policy structure and mechanisms for enforcement that support development of privacy mechanisms for future CSE.

Acknowledgments. This work was partly funded by DFG BO2129/8-1 and the Graduate School IME, Universität Karlsruhe (TH).

References

1. Smyth, B., Balfe, E., Boydell, O.: A Live-User Evaluation of Collaborative Web Search. In: IJCAI (2005)
2. Reddy, M.C., Jansen, B.J., Krishnappa, R.: The Role of Communication in Collaborative Information Searching. In: ASTIS (2008)
3. Morris, M.R., Horvitz, E.: SearchTogether: An Interface for Collaborative Web Search. In: UIST (2007)
4. Fireball (2008), `http://www.fireball.de/`
5. Morris, M.R.: Collaborating Alone and Together: Investigating Persistent and Multi-User Web Search Activities. Technical report, Microsoft Research (2007)
6. EU Data Protection Working Party. Opinion on Data Protection Issues Related to Search Engines (2008)
7. W3.org (2002), `www.w3.org/TR/P3P-preferences/`
8. Cvrcek, D., et al.: A study on the Value of Location Privacy. In: WPES (2006)
9. Burghardt, T., Buchmann, E., Böhm, K.: Discovering the Scope of Privacy Needs in Collaborative Search. In: Web Intelligence (2008)
10. Olson, J.S., Grudin, J., Horvitz, E.: A Study of Preferences for Sharing and Privacy. In: CHI (2005)
11. Acquisti, A., Grossklags, J.: Privacy and Rationality in Individual Decision Making. IEEE Security and Privacy (2005)
12. Consolvo, S., et al.: Location Disclosure to Social Relations: Why, When, & What PeopleWant to Share. In: SIGCHI (2005)
13. Large, A., Beheshti, J., Rahman, T.: Gender Differences in Collaborative Web Searching Behavior: An Elementary School Study. In: Information Processessing and Management (2002)
14. Twidale, M.B., Nichols, D.M., Paice, C.D.: Browsing is a Collaborative Process. In: Information Processing and Management (1997)
15. Glance, N.S.: Community Search Assistant. In: Workshop on AI for Web Search AAAI (2001)

16. Kriewel, S., Fuhr, N.: Adaptive Search Suggestions for Digital Libraries. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 220–229. Springer, Heidelberg (2007)
17. Teevan, J., et al.: Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In: SIGIR (2007)
18. White, R.W., Morris, D.: Investigating the Querying and Browsing Behavior of Advanced Search Engine Users. In: SIGIR (2007)
19. Babbie, E.R.: The Practice of Social Research, 10th edn. Academic Internet Publ., London (2007)
20. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM (1975)
21. Yu, T., Winslett, M.: A Unified Scheme for Resource Protection in Automated Trust Negotiation. In: SP (2003)
22. Zou, D., Liao, Z.: A new Approach for Hiding Policy and Checking Policy Consistency. In: ISA (2008)
23. IPD Privacy Web Site (2008), http://privacy.ipd.uni-karlsruhe.de/
24. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (1995)
25. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, Cambridge (1998)
26. Leskovec, J., Horvitz, E.: Planetary-Scale Views on a Large Instant-Messaging network. In: WWW (2008)
27. Saint-Jean, F., et al.: Private Web Search. In: WPES, Alexandria, Virginia, October 29, pp. 84–90. ACM Press, New York (2007)
28. Bertino, E., Ferrari, E., Squicciarini, A.C.: Trust-x: A peer-to-peer framework for trust establishment. In: TKDE (2004)
29. Vaidya, J., Clifton, C.: Secure set intersection cardinality with application to association rule mining. Journal of Computer Security (2005)
30. Barbaro, M., Zeller Jr., T.: A Face Is Exposed for AOL Searcher No. 4417749 (2006)
31. Samarati, P.: Protecting respondent's privacy in microdata release. In: TKDE (2001)
32. Sweeney, L.: K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems (2002)